Studying biomolecular structures, interactions, and molecular mechanisms using advanced NMR techniques

Konstantinos Tripsianes, Ph.D.

Habilitation

Masaryk University, Brno

Czechia 2026

Contents

Introduction: NMR in Structural Biology ..... p. 2

Structure determination of proteins by NMR ..... p. 5

Streamlining NMR structure determination ..... p. 9

Studying posttranslational modifications by NMR ..... p. 14

Conformational flexibility in biomolecular interactions ..... p. 18

References ..... p. 21

Introduction: NMR in Structural Biology

As a structural tool, Nuclear Magnetic Resonance (NMR) is highly complementary to X-ray crystallography and cryo-electron microscopy (cryo-EM) and uniquely suited for studies of intrinsically disordered and dynamic systems in real time. It produces spectral fingerprints of biomolecules at the atomic scale to provide information on the structure, interactions, and motions that occur in solution. NMR is frequently judged harshly for contributing <10% of the structures in the Protein Data Bank (PDB), yet it is the only method for obtaining high-resolution, atomic-level descriptions of biomolecules in their natural solution milieu.

NMR is an inherently insensitive method and suffers from resonance overlap and broadening that occurs as molecular size increases. These restrictions limit solution NMR studies to relatively small biomolecules, in the case of proteins and nucleic acids to fewer than 500 amino acids or 100 nucleotides, respectively. However, over the years, spectra sensitivity has been improving with the availability of high-field magnets, up to 28.2 Tesla currently (or 1.2 GHz ^{1}H Larmor frequency),[1] the development of specialized probes that minimize electronic noise associated with the detected signals,[2] and the design of new pulse sequences that minimize losses in signal.[3] Other technologies, such as non-uniform sampling,[4] have benefited NMR experiments allowing for greater resolution and signal-to-noise.

NMR spectroscopy uses a strong magnetic field to provide information on the spatial coordinates of nuclei which exhibit the property of 'spin 1/2'. This includes naturally occurring ^{1}H and rarer isotopes ^{13}C and ^{15}N which are introduced into proteins by scientists. Such labeling is routinely achieved by protein expression in *E. coli* with media containing ^{15}N-NH_4Cl and/or ^{13}C-glucose as sole nitrogen and/or carbon sources.[5,6] New isotope labeling schemes are continuously developed that improve both spectral resolution and sensitivity and expand the types of problems that are amenable for investigation.[7,8]

The first step in any NMR study is to assign each peak in the spectrum to the position in the primary structure of the biomolecule from which it derives. This frequently relies on analysis of correlations between nuclei that are connected to each other through bonds.[9] Numerous 3D NMR experiments have been designed that increase spectral resolution and provide the volume of information required to assign the fingerprint spectra of biomolecules.[10] Additional resolution can be obtained through higher dimensionality experiments, such as 4D spectra that correlate four resonance frequencies from different nuclei. However, most of the spectrometer measurement time in a biomolecular NMR project is frequently spent on measuring spectra for the chemical shift assignment, which are often not of direct use to the biological question at stake, such as, for instance, elucidating dynamics or interactions of the protein. The same holds for the time spent by the spectroscopist: Finding chemical shift assignments is time consuming and requires expertise.

Protein structural analysis by NMR relies upon distance restraints obtained from specific NMR experiments, primarily looking at nuclear Overhauser effects (NOE).[11] NOEs show how excited nuclei returning to equilibrium are affected by nearby coupled spins, as a function of distance allowing protein NOESY spectroscopy to determine internuclear distances through space. The collection and analysis of such comprehensive and informative data requires sophisticated technology and significant time and effort from researchers which can be daunting. It has been

the case in the past, that years of work went into resolving just one experimental solution NMR structure.

NMR has a unique ability of providing thermodynamic and structural information on macromolecular complexes with atomic-resolution.[12] Chemical Shift Perturbation (CSP) analysis is probably the most informative and widely applicable NMR method utilized for the extraction of well-defined binding interfaces,[13] detection of diverse affinities (from mM to μM),[14] and the elucidation of ligand binding modes in drug discovery pipelines.[15] In a typical CSP experiment, a reference 2D-heteronuclear single quantum coherence (HSQC) spectrum of a ^{15}N- or ^{13}C-labeled protein is acquired in the absence of its binding partner followed by a series of HSQC spectra measured at increasing concentrations of unlabelled ligand. When the free and bound forms of the labelled protein exchange rapidly on the NMR timescale (i.e., exchange rate ≥ μs^{-1} denoted as 'fast exchange') a progressive change in peak position is observed across the titration. Therefore, a plot of the chemical shift change as a function of the concentration of binding partner results in a binding isotherm that can be fit to obtain the dissociation constant (K_D) for the protein-ligand complex. Mapping CSP at a saturating concentration of binding partner on the structure of the protein provides information on the residues that reside at the interface of the complex. CSP experiments are very useful in comparing the binding mode of different ligands (protein, DNA, or small molecule) to the same target protein.

The development of artificial intelligence (AI) and machine-learning techniques has provided a turning point within protein structure prediction methods. Namely, Google DeepMind's development of AlphaFold, a programme that generates predicted protein structures from amino acid sequences, was awarded the 2024 Nobel Prize in Chemistry for its impact.[16,17] The AlphaFold Protein Structure Database[18] offers immediate free access to predicted structures for almost every known protein, whilst the AlphaFold server[16] allows anyone to predict the structure of a protein or complex of interest, often within minutes. Hence, atomic-level protein structure predictions are now readily accessible to all, and it has rapidly become commonplace to see protein structures in talks by non-structural biologists.

Although a quick read through the many success stories of AlphaFold could create the impression that experimental structure determination is now a thing of the past, this is not completely the case. In most cases, AlphaFold provides reliable structure determinations based on the amino acid sequence of a protein, but experimental structural analyses will remain essential to gain insight into the actual working mechanisms and the interactions of proteins. NMR in particular, is the leading method to study intrinsically disordered proteins (IDPs) or intrinsically disordered regions (IDRs) in biomolecules, which has been estimated to comprise greater than 50% of the human proteome.[19] Thus, the true power of NMR lies in its ability to provide the dynamic view of structural biology, such as real-time monitoring of biochemical reactions, identifying and studying weak or heterogenous interactions made by disordered linkers and IDRs, or mapping ligand binding sites and conformational changes related to function. Interestingly, NMR as a non-destructive and non-invasive technique, has enabled the study of biomolecules inside living cells and holds the potential to yield descriptions of the structures and dynamics within phase-separated compartments.

Despite the power of NMR to study time-dependent processes, motions, and the prospect of applying it in native environments, another serious challenge for NMR is attracting the next

generation of scientists to the discipline. Besides coordinated training programs coupled with improved access to NMR facilities, more user-friendly visualization, analyses, and validation of NMR data could have a tremendous impact in transitioning it to a method widely incorporated into biological research. A key step in this direction is to keep integrating NMR with complementary techniques,[20] such as cryo-EM, X-ray crystallography, mass spectrometry, high-throughput screening, biophysical assays, cellular assays and artificial intelligence-driven modeling, to construct a more comprehensive picture of biomolecular structures, interactions, and mechanisms of function.

My habilitation thesis summarizes the work from different stages of my career, as a PhD student in the Netherlands, as a PostDoc researcher in Germany, and as a Group Leader in Czechia. It is NMR oriented but with a keen interest in biological processes and molecular mechanisms. I wish to thank my mentors, Professor Rolf Boelens and Professor Michael Sattler, my collaborators, many fellows, students, and staff who were critical players in the scientific stories I dicuss below. It is their energy, creativity, and hard work that influenced my scientific career over the past twenty years.

My scientific journey stretches beyond NMR, structures, biological interactions and mechanisms manifested in the papers I make comments on. It is a life journey in different countries and cultures, it is the growing and the support of my family through the years, it is the friendships forged inside or outside the research environment, it is the professional and personal hardships, it is everything that is set aside every time science comes to the fore.

My sincere gratitude goes to my wife Vasia and my kids Eleftheria and Angelakis for putting up with me. Them, it is my journey to Ithaca.

“Keep Ithaca always in your mind.
Arriving there is what you’re destined for.
But don’t hurry the journey at all.
Better if it lasts for years,
so you’re old by the time you reach the island,
wealthy with all you’ve gained on the way,
not expecting Ithaca to make you rich.”

Ithaca by Constantine P. Cavafy

Structure determination of proteins by NMR

At the turn of the century, long before the advent of AlphaFold, a number of structural proteomics initiatives were set up aiming to map out a sizeable fraction of protein fold space.[21,22] The main techniques used were X-ray crystallography and NMR spectroscopy. The advantages of the NMR approach were well known: proteins do not need to be crystallized, experiments are performed in a (close to) physiological environment and are sensitive to protein dynamics. On the down side, however, NMR protein structure calculation was both slow and limited to relatively small targets (~10-30 kDa).

Structural proteomics initiatives were instructive in terms of the ambitions and limits of NMR structural biology. If anything, however, they provided the main impetus for consensus procedures (notably in terms of the optimal set of experiments) and automatic NMR protein structure determination. It should be also mentioned that the developments made were an extension of the two-step protocol established in the 1980s by Wüthrich that survived with few changes.[23] In the first step, the chemical shifts of the NMR-active nuclei in the protein are identified according to their position in the amino-acid sequence. In the second step, a self-constrained picture of the protein fold is modelled through the evaluation of thousands of inter-proton contacts and a biophysical force field.

The NMR structure determination of a protein conventionally involves the preparation of (typically uniformly $^{13}C/^{15}N$-labeled) soluble protein, the acquisition of a set of 2D and 3D NMR experiments, NMR data processing, peak picking, chemical shift assignment, NOE assignment and collection of conformational restraints, structure calculation, refinement and validation.

The chemical shift assignment of proteins is usually performed in four stages: (1) group resonances into spin systems; (2) determine the amino-acid type of each spin system; (3) create fragments from collections of spin systems that match sequentially; (4) associate the said fragments with a particular position in the amino-acid sequence.

The preferred set of 3D experiments that were used in my publications (and still used in lab projects involving intrinsically disordered regions of proteins) for sequence specific backbone assignments are HNCO, HN(CA)CO, HNCA, HN(CA)CB, CBCA(CO)NH, and HBHA(CBCACO)NH.[10] This is the minimum number of through-bond experiments required to establish sequential connectivities based on three carbon atoms (C^{α}, C^{β} and C′), and obtain the full set of backbone chemical shifts (^{15}N, $^{1}H^{N}$, $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$) for empirical prediction of protein backbone φ/ψ torsion angles using TALOS-N.[24] Nowadays, with the use of non-uniform sampling, the six 3D experiments are acquired in less than three days spectrometer time. The assignment task is still performed semi-automated. Several automated procedures have been developed over the years but none of them has been proved error-free. Thus, backbone automated assignments are used as a starting point and curated by the user.

To complete the sequence-specific resonance assignment of a protein the side chain resonances must be assigned as well. Common experiments used for that are the HCCH-COSY and HCCH-TOCSY experiments.[10] These experiments provide through-bond information on the aliphatic side chains which can be linked to the known C^{α}, $^{1}H^{\alpha}$, C^{β}, $^{1}H^{\beta}$

frequencies obtained during the backbone assignment procedure. Obtaining aromatic ^{1}H and ^{13}C assignments in proteins is challenging and is based on the manual analysis of the NOESY spectra.

The classical pathway to NMR structure determination is based on the gathering of nuclear Overhauser effect (NOE) inter-proton distance restraints. The NOE restraints are usually collected from ^{13}C-edited (tuned once for aliphatic and once for aromatic carbons) and ^{15}N-edited 3D NOESY spectra. In order to extract distance restraints from a NOESY spectrum, its cross peaks have to be assigned, i.e. the pairs of interacting hydrogen atoms have to be identified. The assignment of NOESY cross peaks requires as a prerequisite the knowledge of the chemical shifts of the spins from which NOEs are arising, the majority of which is determined earlier by analyzing through-bond experiments. Missing assignments could be still identified by an experienced spectroscopist when interpreting the NOESY spectra manually, to produce lists of chemical shifts of varying completeness and correctness, depending on the quality and information content of the input data.

Obtaining a comprehensive set of distance restraints from a NOESY spectrum is in practice by no means straightforward. Due to the large number of chemical shifts and even larger number of cross peaks in the NOESY spectra, manual NOE assignment is particularly cumbersome and error-prone. The lack of consistency among NMR practitioners and, perhaps laborious interactive procedures that generate bias in NOE interpretation, have led to problematic NMR structures in the early years.[25] The accuracy and quality of NMR structures improved after 2005, mainly due to new concepts implemented in the CYANA algorithm.[26]

CYANA proceeds in iterative cycles of automated NOE assignment followed by structure calculation with torsion angle dynamics. Between subsequent cycles, information is transferred exclusively through the intermediary three-dimensional structures, in that the molecular structure obtained in a given cycle is used to guide the NOE assignments in the following cycle. Otherwise, the same input data are used for all cycles, that is, the amino acid sequence of the protein, one or several chemical shift lists from the sequence-specific resonance assignment, and one or several lists containing the positions and volumes (or intensities) of cross peaks in the NOESY spectra. By considering the iterative nature of CYANA and other algorithms such as ARIA,[27] the success of *de novo* NMR structure determination depends on the ability of the program to find the correct fold in the initial cycle of NOE assignment without any prior knowledge of the structure.

Without going into much detail, three concepts are combined in CYANA to eliminate erroneous assignments at the outset. (1) Ambiguous distance restraints introduced in ARIA to deal with peaks with several assignment possibilities. Ambiguous restraints are always laxer than those corresponding to the native structure but they never distort it as long as the correct assignment is present. (2) Network anchoring to measure how well a particular assignment is supported by the list of all other potential assignments compiled across all available spectra. Correct assignments form a self-consistent set and will score highly, whereas wrong assignments act incoherently in the network and do not support each other. (3) Constraint combination to ensure a significant reduction in the number of artefacts. It involves the implementation of inter-peak ambiguous restraints that will only distort the structure if both peaks are artefacts.

The distance restraints derived from the automated assignment of NOESY spectra are then included in the input for the structure calculation. The second and subsequent cycles differ

from the first cycle by the use of additional selection criteria for NOE assignments that are based on assessments relative to the protein 3D structure from the preceding cycle. The precision of the structure determination normally improves with each subsequent cycle. In the final cycle, an additional filtering step ensures that all NOEs have either unique assignments to a single pair of hydrogen atoms, or are eliminated from the input for the structure calculation. This facilitates the subsequent use of refinement and analysis programs that cannot handle ambiguous distance restraints.[28]

For a structural biologist to be able to pose the right biological questions and design experiments related to protein function, the accuarcy of the NMR structure is of paramount importance. Throughout my career I dealt with many projects and determined twenty five NMR structures using the classical NOE approach. Hands-on experience with CYANA calculations and NMR structure determination in general has taught me that human intervention is required to obtain high-quality structures. The user is in principle able to improve the quality of the input chemical shift and peak lists using the CYANA output which in most cases reveals the problem, e.g. peaks that cannot be assigned and might therefore be artifacts or indications of erroneous or missing sequence-specific assignments. In practice, several runs are performed in input-output feedback modus that may involve additional NOESY experiments to increase the number of distance restraints, manual assignment of missing resonances in the NOESY spectra, curated peak lists deciding which peaks are noise or experimental artifacts and which intensities are distorted by peak overlap or baseline problems, and reasoning as to why NOESY peaks are not assigned or distance restraints are discarded by the algorithm.

In general, a good-quality NMR structure depends on the number of restraints per residue such as NOE-derived distance restraints, dihedral angle restraints predicted form chemical shifts, hydrogen bond restraints, and other geometrical restraints, e.g. disulphide bridges, metal coordination, symmetry. Undoubtedly, the number of long-range distance restraints between methyl and aromatic protons is the critical determinant, as these restraints determine the protein fold. Therefore, special attention must be paid to identify these resonances and if possible, to assign them stereospecifically.[29,30]

Of course, each protein target may pose additional challenges in identifying the resonances, collecting the appropriate number of NOE restraints or implementing non-protein ligands in the structure determination. Based on personal experience (ref. 1-7), most laborious experimental procedures involved homodimers, heterodimers, or protein complexes.

In the case of a symmetric homodimer (ref. 3), the NOESY spectrum contains both intramolecular and intermolecular cross peaks. To unambiguously distinguish intermolecular from intramolecular NOEs, isotope-filtered experiments were recorded on an asymmetrically $^{15}N/^{13}C$-labeled sample.[31] To obtain such a sample, equimolar amounts of uniformly $^{15}N/^{13}C$-labeled and unlabelled proteins were mixed. Random exchange of the subunits yielded 50% of asymmetrically isotope-labelled dimers. Isotope-edited/filtered NMR experiments detected exclusively intermolecular NOEs for this population, whereas no signals were observed for symmetrically labelled or unlabelled homodimers or monomeric species.

For an obligate heterodimer with stability dependence issues (ref. 1), an on-column refolding scheme was devised. The $^{15}N/^{13}C$-labeled subunit (prone to precipitation) was immobilized on the column under denaturing conditions and the unlabelled subunit was circulated in excess over this column while gradually decreasing the urea concentration. The eluted heterodimer

was differentially labeled and isotope-edited/filtered NMR experiments detected exclusively intermolecular NOEs between the subunits, which were mostly masked by the much stronger intramolecular NOEs in the uniformly labelled heterodimer.

For a protein complex the procedure looked much simpler on paper, yet very demanding experimentally (ref. 5). The two proteins were produced in labelled and unlabelled forms and mixed differentially, the stoichiometric complex was isolated, and isotope-edited NOESY experiments (with or without filters) were recorded and combined in structure calculations.

At the end of the day, it is about accuracy, scientific rigor and expected impact of the NMR structure on the biological question. Nowadays, the same issues apply to the reliability of AlphaFold structures. It is true that lengthy experimental procedures have been substituted by few mouse clicks, but do researchers know how to evaluate these structures?

1. **Tripsianes K.**, Folkers G.E., AB E., Das D., Odijk H., Jaspers N.G., Hoeijmakers J.H., Kaptein R., Boelens R.* (2005). The structure of the human ERCC1/XPF interaction domains reveals a complementary role for the two proteins in nucleotide excision repair. *Structure*, *13*(12), 1849-1858. https://doi.org/10.1016/j.str.2005.08.014
2. **Tripsianes K.**, Folkers G.E., Zheng C., Das D., Grinstead J.S., Kaptein R., Boelens R.* (2007). Analysis of the XPA and ssDNA-binding surfaces on the central domain of human ERCC1 reveals evidence for subfunctionalization. *Nucleic Acids Res*,*35*(17), 5789-5798. https://doi.org/10.1093/nar/gkm503
3. Meyer N.H., **Tripsianes K.**, Vincendeau M., Madl T., Kateb F., Brack-Werner R., Sattler M.* (2010). Structural basis for homodimerization of the Src-associated during mitosis, 68-kDa protein (Sam68) Qua1 domain. *J Biol Chem*, *285*(37), 28893-28901. https://doi.org/10.1074/jbc.M110.126185
4. Kateb F., Perrin H., **Tripsianes K.**, Zou P., Spadaccini R., Bottomley M., Franzmann T.M., Buchner J., Ansieau S., Sattler M.* (2013). Structural and functional analysis of the DEAF-1 and BS69 MYND domains. *PLoS One*, *8*(1), e54715. https://doi.org/10.1371/journal.pone.0054715
5. **Tripsianes K.***, Friberg A., Barrandon C., Brooks M., van Tilbeurgh H., Seraphin B., Sattler M.* (2014). A novel protein-protein interaction in the RES (REtention and Splicing) complex. *J Biol Chem, 289*(41), 28640-28650. https://doi.org/10.1074/jbc.M114.592311
6. Anosova I., Melnik S., **Tripsianes K.**, Kateb F., Grummt I.*, Sattler M.* (2015). A novel RNA binding surface of the TAM domain of TIP5/BAZ2A mediates epigenetic regulation of rRNA genes. *Nucleic Acids Res*, *43*(10), 5208-5220. https://doi.org/10.1093/nar/gkv365
7. Meyer N.H., Mayerhofer H., **Tripsianes K.**, Blindow S., Barths D., Mewes A., Weimar T., Kohli T., Bade S., Madl T., Frey A., Haas H., Mueller-Dieckmann J.*, Sattler M.*, Schramm G.* (2015). A Crystallin Fold in the Interleukin-4-inducing Principle of Schistosoma mansoni Eggs (IPSE/alpha-1) Mediates IgE Binding for Antigen-independent Basophil Activation. *J Biol Chem*, *290*(36), 22111-22126. https://doi.org/10.1074/jbc.M115.675066

Streamlining NMR structure determination

As discussed above, conventional NMR structure determination methods rely on a large number of complementary multidimensional datasets and abundant restraint densities (of the order of 15-20 NOE restraints per residue) to obtain reliable, high-resolution structures. Aside from the unavoidable imperfections of experimental NMR spectra, a more fundamental limit is the loss of resolution that follows the protein size as the corresponding slowing of protein motions leads to peak broadening.[11] Perdeuteration is the traditional avenue towards accessing larger proteins and a number of selective labelling schemes have been proposed that promise to extend the range of classical NMR methods.[32] Whereas these labelling schemes, may yield very sparse or, in some cases, no NOE data, they still permit the measurement of chemical shifts, which report on the local magnetic environment of nuclei and offer insights into backbone secondary structure, dynamics, and hydrogen bonding.[33] Therefore it was noted over the years, that chemical shifts can be used to streamline the NMR structure determination process.[34]

The most commonly determined protein shifts are isotropic chemical shifts corresponding to the ^{1}H, ^{13}C and ^{15}N nuclei found in proteins measured in aqueous solution. These kinds of shifts account for more than 99% of all reported chemical shift assignments in the Biological Magnetic Resonance Bank (BMRB). A sufficient body of chemical shift assignments for proteins collected over the years, led to the observation that upfield/downfield trends exist for essentially all backbone $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, $^{1}H^{N}$ and ^{15}N chemical shifts with respect to secondary structure (α-helix or β-strand). This residue-specific capacity of NMR was first exploited with the development of the chemical shift index,[35] a method that uses “digitially filtered” chemical shifts to precisely identify the type and location of protein secondary structures along a protein chain. Since chemical shifts are highly reproducible parameters, they offer truly quantitative structural and dynamic data such as backbone torsion angles as wells as 3D structural coordinates.

It has long been known that chemical shifts are a very sensitive indicator, and hence predictor, of backbone dihedral angles. The fact that backbone resonance assignment is a prerequisite for any protein NMR investigation prompted the development of a number of simple protocols that used chemical shifts to predict backbone torsion angles. However, it was not until the development of TALOS[36] (Torsion Angle Likelihood Obtained from Shift) that the NMR community started using chemical shifts to routinely derive torsion angle restraints. Most torsion angle prediction algorithms, e.g. TALOS-N, DANGLE, and PREDITOR, match sections of a target protein to those from a structural database based on their sequence and chemical shift similarity. The underlying idea is that a particular chemical shift pattern is indicative of a specific set of torsion angles. Thereby, if such a pattern from a query fragment can be matched to one in a database, so the corresponding torsion angles must also agree. The accuracy of these chemical shift-based methods is generally high (80-90% of the torsion angle predictions lie within 30° from corresponding angles in the reference structures),[33] justifying their use in NMR-based protein structure determination in conjunction with NOE restraints.

To address the shortcomings of NOE-based methods, researchers have been exploring various ways to develop chemical-shift-only approaches for determining the 3D structures of proteins. The first successful demonstration dated back in 2001,[37] used a technique known as

chemical shift threading. In protein threading, templates are scored not only by sequence similarity but also by structural and/or physicochemical similarity, such as similarity of secondary structures, accessible surface area, residue hydrophobicity or even chemical shifts. The work involved creating a database of chemical shifts predicted from known PDB structures and then matching the sequence and chemical shifts of the query protein to sequences and chemical shifts of proteins in the database. This technique computed the structure of ubiquitin with good accuracy using only the protein sequence and the corresponding ^{1}H and ^{13}C chemical shifts.

These initial successes in protein structural modelling via chemical shift data energized efforts to further develop this approach and to extend it to fragment-based methods of protein structure determination. While varying in details, fragment-based programs have three main steps where chemical shift data can be used: 1) fragment selection, 2) conformational sampling, and 3) final model selection. The first fragment-based strategy to generate protein structures using only chemical shifts was a program called CHESHIRE.[38] CHESHIRE employs several steps including the prediction of the secondary structure (via chemical shifts and sequence), the generation of torsion angle constraints for well-defined segments, the assembly of the protein fragments or segments into an initial structure followed by the minimization of that structure using a chemical shift force field and a standard MD force field. CHESHIRE was shown to generate high quality structures for 11 query proteins attempted spanning a range of structure classes and sizes (up to 123 residues).

The CS-Rosetta method, developed around the same time as CHESIRE, is based on the same concept.[39] CS-Rosetta combines a highly optimized ab initio fragment assembly protocol,[40] employing different sampling schemes and multiple low-resolution energy functions, together with NMR chemical shifts to yield accurate structural models. First, three- and nine-residue fragments are selected using the backbone chemical shifts, and thus posses φ, ψ backbone torsion angles that are closer to their native values. The low-resolution phase employs a simplified representation of a protein chain, where only backbone heavy atoms and a centroid site representing the sidechain are present for every residue. Thereafter, Monte Carlo fragment trial steps are applied, starting from a fully extended protein chain, to yield collapsed backbone folds. In the high-resolution phase, sidechain atoms are added to low-resolution structures, a rather challenging task due to the exponential number of possible rotamer combinations for all amino acids in a query protein sequence. The final full-atom models are retained based on a scoring function, and quality of fit of the predicted models to the experimental chemical shifts. To further address the conformational sampling problem in order to increase the size limit (>25 kDa), the Resolution Adapted Structural Recombination (RASREC)-Rosetta protocol was developed. In addition to the implementation of an array of optimization strategies, RASREC benefits from parallelization which greatly enhances the sampling range and overall performance of the method. A conceptual breakthrough was to utilize sparse NOE distance restraints to guide RASREC-Rosetta protocol.[41] In particular, RASREC-Rosetta was used together with long-distance NOE data (ILV-methyl protonated samples) recorded for 11 protein targets of sizes up to 40 kDa. These sparse sets of NOE restraints, together with experimental chemical shifts of the backbone atoms, were used by RASREC-Rosetta to model structures at high levels of accuracy. This and many other studies demonstrated that restraints derived from NMR experiments become crucial in modeling protein structures with Rosetta as they guide the conformation search toward a native. It also

suggested that the more accurate modeling provided by the ROSETTA energy function offers an opportunity to further explore NOE information through a more optimal use of chemical shifts in automated NMR structure determination.

My group contributed in developing the autoNOE-Rosetta protocol for simultaneous, automatic NOESY cross-peak assignment and structure generation (ref. 8). This entailed the implementation of a new ROSETTTA module for automatic NOE assignment as well as the development of a robust protocol to couple the iterative search for the near-native protein structures in RASREC with iterative NOE assignment. The integrated approach maximizes the number of “instructive” long-range restraints that can be obtained from the NOE data and ensures the self-consistency of the distance restraints. The method was tested on a benchmark of 50 data sets covering a realistic range of NMR data quality. The NOESY peak lists varied dramatically as a function of the quality of the raw data, the method of picking peaks, and the level of peak list refinement. Similarly, the chemical shift lists were incomplete and, in some cases, contained incorrect assignments. In spite of this wide variety of input data quality, autoNOE-Rosetta yielded accurate results with striking consistency demonstrating a remarkable robustness against challenging input data.

A fortuitous discovery revealed that autoNOE-Rosetta is not only robust against problematic peak lists, but also shows remarkable accuracy in the face of incomplete or erroneous chemical shift assignments. This is a common problem that compromises the quality of NMR structures. During our work on the benchmark, we were puzzled by one small protein traget where autoNOE-Rosetta yielded structures that were 4 A˚ away from the PDB-deposited NMR solution structure (2kj6). This prompted us to investigate whether the structure obtained with autoNOE-Rosetta might actually be better supported by the raw NMR data as well. We carefully reanalyzed the raw input data (NOESY spectra) to find out that the backbone assignment was nearly complete and correct, but the side-chain chemical shifts were incomplete and had miss-assignments. Additionally, the NOESY spectra were under-picked, such that many potentially well resolved peaks were not contained in the original peak list. These issues hindered the structure calculations of NOE-driven programs, but had no influence on the CS-Rosetta calculations. After correcting the input data, the structures obtained with conventional methods matched with the autoNOE-Rosetta models obtained with the incomplete and erroneous data. This exercise showcased that autoNOE-Rosetta is tolerant to 10-15% of errors in sidechain assignments (including methyl group assignments), offering a radical view with regard to fully automation of NMR structure determination. Instead of struglling to utilize a large number (6-10) of complementray NMR spectra that require systematic referencing across them but cannot be relied on to always increase the confidence and coverage of all-atom automated assignments,[42] acknowledge the shortcomings in automated procedures (incompletness and errors in assignments), minimize the input data, and leverage autoNOE-Rosetta robustness in the full NMR pipeline.

With this in mind, my group developed next 4D-CHAINS, an automated procedure to assign backbone and sidechain chemical shifts using only two 4D spectra obtained from a 4D-HCNH TOCSY and a 4D-HCNH NOESY experiment (ref. 9). A third 4D-HCCH experiment is recorded to obtain mainly methyl-methyl NOEs for automated NOE assignment and structure determination with autoNOE-Rosetta.

To address the assignment problem, 4D-CHAINS uses 2D probability density maps of correlated ^{13}C-^{1}H chemical shifts to effectively identify possible spin systems. In particular, 4D-CHAINS combines sequential information present in the 4D-HCNH TOCSY and intraresidue information present in 4D-HCNH NOESY ^{13}C-^{1}H planes, respectively, by clustering TOCSY or NOESY peaks to Amino Acid Index Groups (AAIGs) via their common ^{15}N-^{1}H frequency. Thereafter, AAIGs are mapped to amino acids in the query protein sequence using a procedure similar to genome assembly used in DNA sequencing.[43] During this process, contiguous segments of AAIGs are iteratively matched along the protein sequence until a self-consistent assignment solution is obtained. The 4D-CHAINS protocol was applied and tested to four protein targets ranging from 133 to 248 residues. The algorithm mapped correctly all AAIGs to the respective protein sequences with >95% completeness. The TOCSY-based assignments alone covered approximately 80% of all aliphatic chemical shifts with an error rate of <0.5%. To increase the overall assignment completeness additional information was obtained from the HCNH NOESY spectrum by employing the concept of common NOEs between successive residues.[44] The TOCSY-NOESY combination enabled more complete assignments with 94% correct aliphatic chemical shifts and a combined error rate of 1.3%. It is further worth noting that the vast majority of resonances corresponding to sidechain methyls, which are important probes in identifying the protein fold, were correctly assigned.

The automated 4D-CHAINS assignment lists, which contained errors and unassigned resonances of aromatic, sidechain amide, and plausibly other backbone or sidechain atoms, together with both NOE datasets (HCNH+HCCH) were used as an input for autoNOE-Rosetta calculations. Notably, the ensembles calculated in a in a truly unsupervised manner achieved a high level of convergence and accuracy for all four targets when compared to existing NMR and X-ray structures. Towards evaluating the effect of different levels of assignment completeness on the performance of autoNOE-Rosetta, we carried out benchmark calculations by randomly removing entries from the automated assignment lists and found that autoNOE-Rosetta can identify the correct protein folds from as low as 60–70% sidechain assignments. We also noted that ~25% of the total contacts identified by autoNOE-Rosetta are contributed by methyl NOEs for structure calculations. Overall, our results suggested that the fully automated assignment process introduced by 4D-CHAINS has a minimal impact on the performance and quality of the derived structural ensembles by autoNOE-Rosetta, which remain highly consistent with all available input data.

4D-CHAINS success was based on using a minimal set of 4D spectra to expedite the acquisition time, reduce the NMR experimental cost, minimize the human input required, but most importantly provide higher precision in data analysis and deliver protein structures in less than a month. Thus, our streamlined strategy was able to respond timely to synergistic approaches and collaborative efforts. Using the 4D-CHAINS/autoNOE-Rosetta automated pipeline for resonance assignments and structure calculation we were able to experimentally confirm the first accurate design of an all-β globular protein from scratch by the group of David Baker, US (ref. 10). In another study, we contributed two NMR structures using our protocol in interdisciplinary research led by the group of Daniel van Damme, BE, in understanding the function of EH domains in plant endocytosis (ref. 11). Last but not least, our methods appealed to the group of Vitezlav Bryja, CZ, and were employed to understand the structural effects of phosphomimetic substitutions on a PDZ domain (ref. 12).

Next, we considered to exploit the better NOE build-up and sensitivity in larger proteins and proposed artificial intelligence methods to crack the atomic structures of proteins in a single step. The concept of common NOEs provided the validated rationale for applying artificial intelligence to a set of expected NOE patterns in order to address the assignment problem using only the 4D-NOESY HCNH spectrum. The complexity of the task is prohibitory to human analysis. Instead, deep neural networks were trained on amino acid resonance recognition patterns using the massive amounts of chemical shifts deposited in Biological Magnetic Resonance Bank by human experts (8 million entries) and by taking into account specific features of the 4D-NOESY HCNH data. In addition, simulated HCNH NOESY spectra were generated from VASCO database (~5,000 entries) for high-level training to decompose the intricate NOESY patterns and successfully assign the NOESY spectrum using graph theory. The new software, 4D-GRAPHS, was able to map 80% of residues to correct spin systems, 13% of residues to wrong spin systems, whereas 7% of residues were not mapped, for the four protein targets tested.

When the AlphaFold Protein Structure Database was launched on July 2021, the 4D-GRAPHS project was dropped as there was no longer need for high-throughput structural determination.

8. Zhang Z., Porter J., **Tripsianes K.**, Lange O.F.* (2014). Robust and highly accurate automatic NOESY assignment and structure determination with Rosetta. *J Biomol NMR*, *59*(3), 135-145. https://doi.org/10.1007/s10858-014-9832-4

9. Evangelidis, T., Nerli, S., Novacek, J., Brereton, A. E., Karplus, P. A., Dotas, R. R., Venditti, V., Sgourakis, N. G.*, & **Tripsianes, K.*** (2018). Automated NMR resonance assignments and structure determination using a minimal set of 4D spectra. *Nat Commun*, *9*(1), 384. https://doi.org/10.1038/s41467-017-02592-z

10. Marcos, E.*, Chidyausiku, T. M., McShan, A. C., Evangelidis, T., Nerli, S., Carter, L., Nivon, L. G., Davis, A., Oberdorfer, G., **Tripsianes, K.**, Sgourakis, N. G., & Baker, D.* (2018). De novo design of a non-local beta-sheet protein with high stability and accuracy. *Nat Struct Mol Biol*, *25*(11), 1028-1034. https://doi.org/10.1038/s41594-018-0141-6

11. Harnos, J., Canizal, M. C. A., Jurasek, M., Kumar, J., Holler, C., Schambony, A., Hanakova, K., Bernatik, O., Zdrahal, Z., Gomoryova, K., Gybel, T., Radaszkiewicz, T. W., Kravec, M., Trantirek, L., Rynes, J., Dave, Z., Fernandez-Llamazares, A. I., Vacha, R., **Tripsianes, K.**, Hoffmann, C., Bryja, V.* (2019). Dishevelled-3 conformation dynamics analyzed by FRET-based biosensors reveals a key role of casein kinase 1. *Nat Commun*, *10*(1), 1804. https://doi.org/10.1038/s41467-019-09651-7

12. Yperman, K., Papageorgiou, A. C., Merceron, R., De Munck, S., Bloch, Y., Eeckhout, D., Jiang, Q., Tack, P., Grigoryan, R., Evangelidis, T., Van Leene, J., Vincze, L., Vandenabeele, P., Vanhaecke, F., Potocky, M., De Jaeger, G., Savvides, S. N.*, **Tripsianes, K.***, Pleskot, R.*, & Van Damme, D.* (2021). Distinct EH domains of the endocytic TPLATE complex confer lipid and protein binding. *Nat Commun*, *12*(1), 3050. https://doi.org/10.1038/s41467-021-23314-6

Studying posttranslational modifications by NMR

Post-translational modifications (PTMs) are covalent alterations of polypeptide chains that provide a mechanism for expanding the cellular proteome after protein translation is complete.[45] There are over 200 chemical moieties that are enzymatically conjugated to 15 of the 20 naturally occurring amino acids in a protein. PTMs vary considerably in complexity and size of the conjugated group, ranging from the simple oxidation of Cys residues to form disulfide linkages, to the attachment of heavy polyubiquitin or oligosaccharide units. Amino acid side chains carrying a nucleophilic functional group, such as the hydroxyl group of Ser and Thr, the thiol in Cys, and the amine variants of Lys and Arg are most susceptible to modification.

PTMs are not templated but are contingent upon stochastic recognition and modification by enzymes. In addition, PTMs are generally reversible: for example, amino acid-specific kinases add phosphate groups to Ser, Thr, His, or Tyr residues, which are subsequently removed by the action of phosphatases. The relative activities of the enzymes responsible for the addition and removal of the modification also play a role in controlling the extent of PTM at a particular protein site.[46]

PTMs typically occur in regulatory protein regions that are intrinsically disordered and expand the functional repertoire of the polypeptide sequence by altering physicochemical properties such as charge, hydrophobicity and/or conformation. The amino acid sequence of the target protein usually encodes the specificity determinants for the modifying enzymes and for the protein modules that eventually recognize the different PTMs.[47] In this respect it is known that PTMs modulate protein-protein interactions, but in particular cases, they can also trigger structural rearrangements that, in turn, differentially regulate functional, biological outcomes. Thus, from the NMR perspective, one line or research has focused on mapping protein PTM sites *in vitro* or in complex environments such as cell extracts and intact cells,[48] and another line on addressing structure-function relationships brought about by PTMs (ref. 13, 15, 18).

PTM detection by NMR spectroscopy is subject to the same inherent limitations as all other high-resolution NMR applications. Increasing protein/PTM-substrate sizes inevitably lead to greater spectral complexities and unfavorable NMR relaxation behavior. Residue-resolved PTM site mapping requires isotope labelling ($^{13}C/^{15}N$), triple-resonance NMR experiments and dedicated NMR backbone assignment routines. These routines must be usually performed twice, once for the unmodified version and once for the PTM-version of the protein. Provided that resonance-specific assignments are available, the corresponding PTM site(s) are mapped using 2D ^{1}H-^{15}N and ^{1}H-^{13}C heteronuclear correlation methods.[49] Experiments of that sort afford higher resolution insights into PTM reactions and provide residue-resolved, positional information about PTM target sites.

NMR spectroscopy additionally offers convenient means for time-resolved NMR measurements of reconstituted PTM reactions *in vitro*,[50] but also of cellular modification events in complex environments such as cell extracts.[51] The quantitative nature of NMR spectroscopy makes it particularly appealing for PTM studies. Because changes in NMR signal intensities of modified and unmodified substrate residues are detected side-by-side, substrate/product concentrations, and time-dependent changes thereof, are readily deduced and kinetic reaction parameters can directly be extracted.[52] Such measurements are particularly useful in providing

additional mechanistic insights into stepwise PTM reactions that require completions of certain PTM events, before others can ensue.[53]

Structural characterization of PTMs by NMR possesses additional challenges as the PTMs are usually not available in an isotope-labeled form. This requires NMR labeling schemes and tailored NMR experiments for chemical shift assignment of PTMs and strategies to overcome problems with signal overlap and protein-PTM NOE ambiguities (ref. 16). Of course, the overall success depends on additional factors unique to the system under investigation, e.g. strength of the interaction, exchange processes induced by binding, number of PTM chemical shifts and their degeneracy. Depending on the complexity of the system, individual steps were adapted and the workflow was supplemented accordingly to study a spectrum of PTMs including arginine methylation (ref. 13), C-terminal amidation (ref. 14), farnesylation (ref. 15), phosphorylation (ref. 17), and C-terminal polyglutamylation (ref. 18). To decipher the readout of arginine dimethylation and the conformational changes induced by farnesylation, it required methodological advancements for the collection of NOE restraints and NMR structure determination.

Arginine methylation plays pivotal roles in the formation of complexes that are involved in a variety of cellular processes, including RNA processing, transcriptional regulation, trafficking, signal transduction and gametogenesis. Our studies have shed light on the structural and energetic determinants of dimethylarginine recognition by the Tudor domains of SMN and SPF30 (ref. 13). The prototypic Tudor domain is a segment of approximately 60 amino acids which forms a β-barrel like core structure and recognizes dimethylated arginine (DMA) residues. Protein arginine methyltransferases generate asymmetric DMA (aDMA) by placing two methyl groups on one of the terminal guanidino nitrogen atoms of arginine, or symmetric DMA (sDMA) by placing one methyl group on each of the terminal guanidino nitrogen atoms, whereas monomethylated arginine (MMA) is probably generated as an intermediate *en route* to dimethylation of target proteins.

Binding assays using isothermal titration calorimetry demonstrated that the SMN and SPF30 Tudor domains do not bind arginine, as well as lysine or its (mono-, di- and tri-) methylated forms. Arginine methylation, on the other hand, induced measurable affinities, with binding increasing from MMA to aDMA to sDMA. Yet the measured affinities were in the low millimolar range (~0.5 mM) precluding, at first sight, structural investigations that require high level of protein saturation in complex with DMAs. Saturation is determined by the binding affinity and to achieve it at a low protein-DMA ratio the concentration of both components must exceed largely the determined affinity. Indeed, NMR titration experiments using 200 μM of the Tudor domain and up to 20 molar excess of the ligand (4 mM) resulted in ~75% saturation of the Tudor domain. When the same experiments were performed using 4 mM Tudor domain and up to 4.5 molar excess of the ligand (18 mM), Tudor saturation exceeded 95%. This breakthrough circumvented previous technical restrictions that precluded structural analysis,[54,55] but most importantly alleviated sensitivity issues associated with isotope-filtered and isotope-edited 3D NOESY experiments due to the high concentration of the Tudor-DMA complexes.

Four high-quality structures were calculated, based on a large number of total (more than 1700) and intermolecular (more than 70) distance restraints for each of the SMN-sDMA, SPF30-sDMA, SMN-aDMA, and SPF30-aDMA complexes. The distance restraints were

obtained from a total of 24 3D NOESY experiments, six for each structure, three without and three with $^{15}N/^{13}C$ filters in the indirect F1 dimension. For CYANA calulations the DMA libraries were prepared manually using the PDB coordinates derived from the PRODRG2 Server and for water refinement the parameter and topology files were taken and adapted from the PRODRG2 Server as well.

The NMR structures showed that the same general principles allow the Tudor domains of SMN and SPF30 to bind aDMA and sDMA. In either complex, the planar guanidino group of the DMA is recognized by a Tudor cage formed by four aromatic residues. Arginine dimethylation imparts hydrophobicity and bulkiness to the residue, but does not neutralize the cationic charge. As a common principle of DMA recognition, the steric bulk provided by the addition of two methyl groups to the arginine side chain guides the position of the cation relative to the rigid aromatic cage to efficiently utilize cation-π interactions in an otherwise hydrophobic environment. This mode of recognition achieves a specific readout of the dimethylation marks and discriminates against non- and mono-methylated arginines.

Additional binding experiments showed that Tudors are able to achieve increased affinities for arginine- and glycine-rich peptides by external binding cooperativity due to the presence of multiple DMA modifications, provided a functional link to Tudor missense mutations associated with Spinal Muscular Atrophy, demonstrated the role of the aromatic cage in DMA recognition, and predicted that about half of the Tudor domains in the TDRD family and the TUDOR *Drosophila* protein may not support efficient recognition of the dimethyl-arginine modification.

Farnesylation is a prenylation process wherein a 15-carbon farnesyl group is covalently attached to the cysteine residue of target proteins by farnesyltransferases. It is an irreversible posttranslational modification mostly found in small GTPases and is known to mediate protein-protein and membrane-protein associations. Our studies on PEX19, a peroxisomal biogenesis factor required for the transport of peroxisomal membrane proteins (PMPs), revealed an allosteric mechanism for the modulation of protein function by farnesylation (ref. 15). Farnesylation can be performed *in vitro* using purified PEX19, the farnesyl transferase enzyme, farnesyl pyrophosphate as the donor of (2*E*,6*E*)-farnesol, and cofactors such as Mg^{2+} and Zn^{2+} to obtain fully farnesylated PEX19 protein. As regards NMR structural investigations, PEX19 labeling can vary but the farnesyl is always at natural isotopic abundance.

To identify the farnesyl resonances, we prepared perdeuterated PEX19 (97% ^{2}H) without ^{13}C labeling to reduce the occurrence of ^{13}C-bound protons in the protein (at natural abundance) to 0.03%. With this labeling scheme the farnesyl resonances were readily detected at natural abundance (1%) without interference from protein signals, allowing to unambiguously identify the chemical shifts and deduce the stereochemistry of the farnesyl moiety when attached to the protein. PEX19 perdeuteration was next combined in a series of different samples with amino acid specific labeling ($^{1}H/^{15}N/^{13}C$ leucine; $^{1}H/^{15}N/^{13}C$ isoleucine; $^{1}H/^{13}C\varepsilon$ methionine), reverse labeling of phenylalanines, and selective methyl labeling of isoleucine, leucine, and valine, to complete and resolve ambiguities of PEX19 assignments but also to obtain protein-farnesyl NOE correlations, which would otherwise be masked by protein-protein correlations in the NOESY spectra. The PEX19 farnesylated structure was determined in a semi-automated and iterative procedure where additional peaks were manually asssigned to account for distance restraints involving the farnesyl atoms. The solution structure of the farnesylated PEX19 was determined based on numerous NOE-derived distance restraints

including 203 protein-farnesyl restraints, to reveal a bent conformation of the lipid moiety buried inside a hydrophobic cavity of the protein.

The combination of the various labeling protocols with individually adapted NMR experiments enabled us to circumvent the restrictions for the unlabeled ligand, to obtain complete resonance assignments for both protein and the attached lipid and to determine the first NMR structure of a farnesylated protein.

To our knowledge, this was the first example of a protein, which accommodates the farnesyl group by an internal hydrophobic cavity formed by the same polypeptide chain *in cis*. This induces substantial conformational changes that allosterically reshape the PEX19 surface to form two hydrophobic pockets for the recognition of conserved aromatic/aliphatic side chains in PMPs. Mutations of PEX19 residues that either mediate farnesyl contacts or are directly involved in PMP recognition abolish cargo binding and cannot complement a *Δ*PEX19 phenotype in human Zellweger patient fibroblasts.

13. **Tripsianes K.**, Madl T., Machyna M., Fessas D., Englbrecht C., Fischer U., Neugebauer K.M., Sattler M.* (2011). Structural basis for dimethylarginine recognition by the Tudor domains of human SMN and SPF30 proteins. *Nat Struct Mol Biol*,*18*(12), 1414-1420. https://doi.org/10.1038/nsmb.2185

14. Rodriguez Camargo D.C., **Tripsianes K.**, Buday K., Franko A., Gobl C., Hartlmuller C., Sarkar R., Aichler M., Mettenleiter G., Schulz M., Boddrich A., Erck C., Martens H., Walch A.K., Madl T., Wanker E.E., Conrad M., de Angelis M.H., Reif B.* (2017). The redox environment triggers conformational changes and aggregation of hIAPP in Type II Diabetes. *Sci Rep*, *7*, 44041. https://doi.org/10.1038/srep44041

15. Emmanouilidis L., Schütz U., **Tripsianes K.***, Madl T., Radke J., Rucktaschel R., Wilmanns M., Schliebs W., Erdmann R., Sattler M.* (2017). Allosteric modulation of peroxisomal membrane protein recognition by farnesylation of the peroxisomal import receptor PEX19. *Nat Commun*, *8*, 14635. https://doi.org/10.1038/ncomms14635

16. **Tripsianes, K.***, Schutz, U., Emmanouilidis, L., Gemmecker, G., & Sattler, M.* (2019). Selective isotope labeling for NMR structure determination of proteins in complex with unlabeled ligands. *J Biomol NMR*, *73*(3-4), 183-189. https://doi.org/10.1007/s10858-019-00241-9

17. Hanakova, K., Bernatik, O., Kravec, M., Micka, M., Kumar, J., Harnos, J., Ovesna, P., Paclikova, P., Radsetoulal, M., Potesil, D., **Tripsianes, K.**, Cajanek, L., Zdrahal, Z., & Bryja, V.* (2019). Comparative phosphorylation map of Dishevelled 3 links phosphosignatures to biological outputs. *Cell Commun Signal*, *17*(1), 170. https://doi.org/10.1186/s12964-019-0470-z

18. Kravec, M., Sedo, O., Nedvedova, J., Micka, M., Sulcova, M., Zezula, N., Gomoryova, K., Potesil, D., Sri Ganji, R., Bologna, S., Cervenka, I., Zdrahal, Z., Harnos, J., **Tripsianes, K.**, Janke, C., Barinka, C., & Bryja, V.* (2024). Carboxy-terminal polyglutamylation regulates signaling and phase separation of the Dishevelled protein. *EMBO J*, *43*(22), 5635-5666. https://doi.org/10.1038/s44318-024-00254-7

Conformational flexibility in biomolecular interactions

It is estimated that approximately 30% of eukaryotic proteomes consists of intrinsically disordered regions (IDRs).[56] Despite lacking a fixed structure, IDRs are essential for cellular processes such as replication,[57] transcription,[58] signalling,[59] and many others. At the molecular level, IDRs can function as flexible linkers, as tunable modules for molecular recognition, as binding interfaces for simultaneous interactions with multiple partners, and as drivers of subcellular organization.[60] A common feature shared by many IDRs in different functions is their ability to enable multivalent, tunable and malleable molecular recognition.

Instead of a stable 3D structure, IDRs exist in a collection of rapidly interconverting structurally distinct conformations known as an ensemble. Just as structure describes a folded domain, an IDR can be quantitatively described by its ensemble properties e.g. global dimensions, local transient structure, and inter-residue distances, described in terms of probability distributions. Understanding that IDRs are defined by sequence-specific ensembles with unique physicochemical features acknowledges that their properties can alter in response to molecular interactions, changes in the cellular environment or post-translational modifications.

The last years, the role of IDRs in biomolecular phase transitions has captured increasing attention. Assemblies formed via phase transitions are often called biomolecular condensates, defining membraneless compartments in cells that concentrate specific biomolecules and exclude others.[61] One reason why IDRs are often found to be associated with phase transitions is the multivalent nature of their interactions.[62]

NMR is the primary tool in characterizing IDRs and their functional interactions, among other biophysical methods such as single-molecule Förster resonance energy transfer (FRET), small-angle X-ray scattering (SAXS), and hydrogen–deuterium exchange mass spectrometry (HDX-MS). One of the major hurdles is the NMR characterization of proteins containing both ordered and disordered regions. This poses major challenges, because resonances arising from residues in the structured domain(s) experience substantial broadening while those associated with the disordered domain(s) have narrow linewidths but may be poorly resolved due to limited spectral dispersion and extensive resonance overlap. Approaches to elucidate the structure and dynamics of isolated fragments comprising either disordered or structured regions do not provide the vital insights into the physiological context. For IDRs connected to folded domains, the steric impact of the folded regions and their chemical makeup can significantly influence IDR ensemble properties and function. PTMs offer another way to alter IDR physiological context. PTMs enable covalent but reversible changes in IDR sequence chemistry, which can influence intramolecular and intermolecular interactions.

My group at CEITEC MU is using NMR spectroscopy and complementary techniques in capturing synergies between ordered and disordered regions, intramolecular or intermolecular, and deciphering the underlying molecular mechanisms. The research statement is conceptualized as follows: “Neither disordered regions nor structured domains are independent of one another, but sense and respond to each other and have co-evolved to have some synergistic properties that are required to enable key biological functions. Moreover, protein conformational plasticity mediated by order-disorder interactions should allow the combinatorial integration of phosphorylation (or other) input, partners binding, assembly in droplets, and/or allosteric coupling, to exquisitely control context-dependent

functions". Below, I highlight three recent studies of my group regarding one IDR that switches between phases, conformational states and binding partners (ref. 19), another IDR that stays bound to its partner as it constantly alternates between two binding modes (ref. 20), and a third IDR with a switch-like behaviour in response to multiphosphorylation (ref. 21).

Helicases are ubiquitous molecular motor proteins involved in all aspects of DNA metabolism.[63] Their primary function is to catalyse duplex DNA unwinding in an ATP-dependent fashion.[64] A subset of DNA helicases is also known to unwind non-canonical secondary DNA structures, called G-quadruplexes (G4s),[65] which induce genome instability if left unresolved. RECQ4 protein is a helicase implicated in G4 metabolism,[66] which contains an intrinsically disordered region (IDR) at the N-terminus of its sequence.

Biophysical analysis uncovered a positively charged polyelectrolyte in the IDR of RECQ4 helicase. The structural, kinetic, and thermodynamic profile of its interactions revealed distinct modes of binding depending on the partner. On one hand, the IDR remains highly flexible in polyelectrolyte interactions with nucleic acids. The disordered state in these complexes minimizes the entropic penalty of binding, resulting in relatively strong but non-specific interactions. On the other hand, the IDR forms a stable helix when it interacts with replication protein A (RPA), the major protein that binds to single-stranded DNA in eukaryotic cells. The induced folding comes at an entropic cost, resulting in an interaction characterized by high specificity but low affinity. NMR competition experiments show that the two binding modes are mutually exclusive.

Interestingly, when the RECQ4 IDR binds to G-quadruplexes, liquid-liquid phase separation (LLPS) occurs. By integrating equilibrium and kinetic binding data in a global numerical model, a three-step model of charge-driven coacervation was deduced. The oppositely charged IDR and G4 molecules form a complex in the solution that follows a rapid nucleation-growth mechanism leading to a dynamic equilibrium between dilute and condensed phases. Additional experiments using differential labeling of the interacting partners confirmed the exchange between the solution and droplets. The dynamic nature of coacervation and the ability to explore the two extremes of the structural continuum of complexes in selecting partners (RPA or DNA) suggest a regulatory role for the IDR to enable diverse RECQ4 functions in DNA replication and DNA repair.

The central component of Wnt signaling is the cytoplasmic protein Dishevelled (DVL).[67] DVL is a modular protein made up of three globular domains (DIX, PDZ, DEP) connected by intrinsically disordered regions and a long C-terminal disordered tail (IDRs). DVL function is regulated by a PDZ-ligand interaction in *cis*. The DVL C-terminal tail is tucked away in its PDZ pocket, acting as a DVL conformational switch that directs different biological outputs.[68]

Driven by our quest to understand the multifaceted function of DVL in Wnt signaling, we set out to mechanistically characterize the DVL conformational switch. The crystal structure of the DVL3 PDZ domain in complex with the class III ligand derived from its own C-terminus highlighted a dual binding mode. The PDZ domain is able to recognize either the C-terminal main-chain carboxylate or the internal side-chain carboxylate of a fully conserved aspartate. The two binding modes are virtually indistinguishable due to two registers that fulfill highly similar binding features. Single-carboxylate variants were sufficient to anchor the ligand in the PDZ, in support of the crystallographic observations. Relaxation dispersion experiments of PDZ-ligand fusion proteins demonstrated that the PDZ domain was in exchange between two

states only when in complex with the native class III ligand. Further NMR analysis demonstrated that the class III ligand alternates between terminal and internal binding modes in an equilibrium of nearly equal populations as it engages with the PDZ domain. This study expanded the structural portrait of PDZ interactions in an unprecedented fashion, as it described the PDZ-ligand dynamics at the binding interface.

The current model of DVL function in the Wnt/β-catenin pathway suggests that upon binding of Wnt to its receptor Frizzled (FZD), DVL is recruited to the plasma membrane, where it directly interacts with FZD through its DEP domain.[69,70] However, how Wnt-induced phosphorylation of DVL by Casein kinase 1 (CK1) propagates downstream signaling is an open question and the least understood step in the Wnt signaling pathways.

We identified phosphorylation sites in endogenous DVL induced by Wnt ligands under physiological conditions and addressed the consequences of phosphorylation on DVL conformation, function and signaling. Our analysis pointed to an S/T cluster of IDR2 proximal to the DEP domain that is heavily phosphorylated by CK1. The accumulation of negative charge in IDR2 enhances intramolecular binding to the DEP domain and drives a switch-like response that alters DVL3 conformation when a threshold level of phosphorylation is reached. This conformational change is incompatible at the molecular level with DVL binding to FZD receptors. Indeed, proximity interactomics of DVL mutants revealed FZD receptors as the prominent effectors of the DVL phospho-switch function consistent with the molecular competition at the DEP interface. By integrating findings at different levels, we proposed a mechanism, in which Wnt-induced DVL conformational phospho-switch attenuates co-residence with FZDs as a means for downstream signaling events.

It remains to be determined how conserved this mechanism is. Sequence differences exist among DVLs and binding partners across individual paralogs and species. Our future studies aim at uncovering DVL paralog-specfic mechanistic secrets in Wnt signaling. Our research combines bioinformatics, AlphaFold3 predictions, rigorous NMR anaysis, biochemical and biophysical assays, in combination with cell-based assays, interactomics and animal models.

19. Papageorgiou, A. C., Pospisilova, M., Cibulka, J., Ashraf, R., Waudby, C. A., Kaderavek, P., Maroz, V., Kubicek, K., Prokop, Z., Krejci, L.*, & **Tripsianes, K.*** (2023). Recognition and coacervation of G-quadruplexes by a multifunctional disordered region in RECQ4 helicase. *Nat Commun*, *14*(1), 6751. https://doi.org/10.1038/s41467-023-42503-z

20. Kumar, J., Micka, M., Komarek, J., Klumpler, T., Bystry, V., Sprangers, R., Barinka, C., Bryja, V., & **Tripsianes, K.*** (2025). A class III ligand oscillates between internal and terminal binding modes as it engages with the Dishevelled PDZ domain. *Structure*, *33*(8), 1362-1373 https://doi.org/10.1016/j.str.2025.05.012

21. Micka, M., Kumar, J., Paclíková, P., Hayek, Z., Hanáková, K., Bechara, C., Plešingerová, H., Šedo, O., Bologna, S., Del Nero, E., Gömöryová, K., Bystrý, V., Gybeľ, T., Číhalová, T., Kravec, M., Potěšil, D., Zdráhal, Z., **Tripsianes, K.*** & Bryja, V.* (2026). A Wnt-induced conformational phospho-switch in DVL3 controls association with Frizzled receptors and Wnt/β-catenin signaling. *Science Advances* (in revision)

References

1. Patching, S. G. Ultra-High Field Nuclear Magnetic Resonance (NMR) Spectroscopy: Applications at 1 GHz and Beyond. *Int J Curr Sci Res Rev* **8**, 336-351, (2025).

2. Li, Q. *et al.* Applications of high-field nuclear magnetic resonance (NMR) in chemical research. *Journal of Magnetic Resonance Open* **23**, 100199, (2025).

3. Joseph, D. *et al.* Optimal control pulses for the 1.2-GHz (28.2-T) NMR spectrometers. *Sci Adv* **9**, eadj1133, (2023).

4. Robson, S. *et al.* Nonuniform Sampling for NMR Spectroscopy. *Methods Enzymol* **614**, 263-291, (2019).

5. Jansson, M. *et al.* High-level production of uniformly ^{15}N- and ^{13}C-enriched fusion proteins in Escherichia coli. *J Biomol NMR* **7**, 131-141, (1996).

6. Cai, M. *et al.* An efficient and cost-effective isotope labeling protocol for proteins expressed in Escherichia coli. *J Biomol NMR* **11**, 97-102, (1998).

7. Lacabanne, D. *et al.* Selective labeling and unlabeling strategies in protein solid-state NMR spectroscopy. *J Biomol NMR* **71**, 141-150, (2018).

8. Rowlinson, B. *et al.* Specific isotopic labelling and reverse labelling for protein NMR spectroscopy: using metabolic precursors in sample preparation. *Biochem Soc Trans* **50**, 1555-1567, (2022).

9. Cavanagh, J. *Protein NMR spectroscopy : principles and practice*. 2nd edn, (Academic Press, 2007).

10. Sattler, M. *et al.* Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Progress in NMR Spectroscopy* **34**, 93-158, (1999).

11. Neuhaus, D. Nuclear Overhauser Effect. *eMagRes*, (2011).

12. Purslow, J. A. *et al.* NMR Methods for Structural Characterization of Protein-Protein Complexes. *Front Mol Biosci* **7**, 9, (2020).

13. Mureddu, L. *et al.* Simple high-resolution NMR spectroscopy as a tool in molecular biology. *FEBS J* **286**, 2035-2042, (2019).

14. Williamson, M. P. Using chemical shift perturbation to characterise ligand binding. *Prog Nucl Magn Reson Spectrosc* **73**, 1-16, (2013).

15. Li, Q. *et al.* A Practical Perspective on the Roles of Solution NMR Spectroscopy in Drug Discovery. *Molecules* **25**, (2020).

16. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589, (2021).

17. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590-596, (2021).

18. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* **50**, D439-D444, (2022).

19. Dyson, H. J. *et al.* NMR illuminates intrinsic disorder. *Curr Opin Struct Biol* **70**, 44-52, (2021).

20. Schwalbe, H. *et al.* The future of integrated structural biology. *Structure* **32**, 1563-1580, (2024).

21. Levitt, M. Nature of the protein universe. *Proc Natl Acad Sci U S A* **106**, 11079-11084, (2009).

22. Terwilliger, T. C. *et al.* Lessons from structural genomics. *Annu Rev Biophys* **38**, 371-383, (2009).

23. Wüthrich, K. NMR with Proteins and Nucleic Acids. *Europhys. News* **17**, 11-13, (1986).

24. Shen, Y. *et al.* Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J Biomol NMR* **56**, 227-241, (2013).

25. Fowler, N. J. *et al.* The accuracy of NMR protein structures in the Protein Data Bank. *Structure* **29**, 1430-1439 e1432, (2021).

26. Guntert, P. Automated structure determination from NMR spectra. *Eur Biophys J* **38**, 129-143, (2009).

27. Linge, J. P. *et al.* ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* **19**, 315-316, (2003).

28. Nederveen, A. J. *et al.* RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. *Proteins* **59**, 662-672, (2005).

29. Neri, D. *et al.* Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 repressor by biosynthetically directed fractional 13C labeling. *Biochemistry* **28**, 7510-7516, (1989).

30. Kainosho, M. *et al.* Optimal isotope labelling for NMR protein structure determinations. *Nature* **440**, 52-57, (2006).

31. Breeze, A. L. Isotope-filtered NMR methods for the study of biomolecular structure and interactions. *Prog Nucl Mag Res Sp* **36**, 323-372, (2000).

32. Ohki, S.-y. *et al.* Stable isotope labeling methods for protein NMR spectroscopy. *Prog Nucl Mag Res Sp* **53**, 208-226, (2008).

33. Wishart, D. S. Interpreting protein chemical shift data. *Prog Nucl Magn Reson Spectrosc* **58**, 62-87, (2011).

34. Nerli, S. *et al.* Chemical shift-based methods in NMR structure determination. *Prog Nucl Magn Reson Spectrosc* **106-107**, 1-25, (2018).

35. Wishart, D. S. *et al.* The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data. *J Biomol NMR* **4**, 171-180, (1994).

36. Cornilescu, G. *et al.* Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR* **13**, 289-302, (1999).

37. Wishart, D. S. *et al.* Use of chemical shifts in macromolecular structure determination. *Methods Enzymol* **338**, 3-34, (2001).

38. Cavalli, A. *et al.* Protein structure determination from NMR chemical shifts. *Proc Natl Acad Sci U S A* **104**, 9615-9620, (2007).

39. Shen, Y. *et al.* Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci U S A* **105**, 4685-4690, (2008).

40. Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* **487**, 545-574, (2011).

41. Lange, O. F. *et al.* Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proc Natl Acad Sci U S A* **109**, 10873-10878, (2012).

42. Schmidt, E. *et al.* A new algorithm for reliable and general NMR resonance assignment. *J Am Chem Soc* **134**, 12817-12829, (2012).

43. Li, Z. *et al.* Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief Funct Genomics* **11**, 25-37, (2012).

44. Xu, Y. *et al.* A new strategy for structure determination of large proteins in solution without deuteration. *Nat Methods* **3**, 931-937, (2006).

45. Walsh, C. T. *Posttranslational modification of proteins: Expanding nature's inventory*. (Roberts & Company Publishers, 2005).

46. Lin, H. *et al.* Introduction: Posttranslational Protein Modification. *Chem Rev* **118**, 887-888, (2018).

47. Seet, B. T. *et al.* Reading protein modifications with interaction domains. *Nat Rev Mol Cell Biol* **7**, 473-483, (2006).

48. Theillet, F. X. *et al.* Cell signaling, post-translational protein modifications and NMR spectroscopy. *J Biomol NMR* **54**, 217-236, (2012).

49. Kumar, A. *et al.* Characterizing Post-Translational Modifications and Their Effects on Protein Conformation Using NMR Spectroscopy. *Biochemistry* **59**, 57-73, (2020).

50. Theillet, F. X. *et al.* Site-specific NMR mapping and time-resolved monitoring of serine and threonine phosphorylation in reconstituted kinase reactions and mammalian cell extracts. *Nat Protoc* **8**, 1416-1432, (2013).

51. Selenko, P. *et al.* In situ observation of protein phosphorylation by high-resolution NMR spectroscopy. *Nat Struct Mol Biol* **15**, 321-329, (2008).

52. Cordier, F. *et al.* Quantitative and dynamic analysis of PTEN phosphorylation by NMR. *Methods* **77-78**, 82-91, (2015).

53. Theillet, F. X. *et al.* Site-specific mapping and time-resolved monitoring of lysine methylation by high-resolution NMR spectroscopy. *J Am Chem Soc* **134**, 7616-7619, (2012).

54. Selenko, P. *et al.* SMN tudor domain structure and its interaction with the Sm proteins. *Nat Struct Biol* **8**, 27-31, (2001).

55. Sprangers, R. *et al.* High-resolution X-ray and NMR structures of the SMN Tudor domain: conformational variation in the binding site for symmetrically dimethylated arginine residues. *J Mol Biol* **327**, 507-520, (2003).

56. Holehouse, A. S. *et al.* The molecular basis for cellular function of intrinsically disordered protein regions. *Nat Rev Mol Cell Biol* **25**, 187-211, (2024).

57. Parker, M. W. *et al.* A new class of disordered elements controls DNA replication through initiator self-assembly. *Elife* **8**, (2019).

58. Brodsky, S. *et al.* Order through disorder: The role of intrinsically disordered regions in transcription factor binding specificity. *Curr Opin Struct Biol* **71**, 110-115, (2021).

59. Wright, P. E. *et al.* Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol* **16**, 18-29, (2015).

60. van der Lee, R. *et al.* Classification of intrinsically disordered regions and proteins. *Chem Rev* **114**, 6589-6631, (2014).

61. Banani, S. F. *et al.* Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol* **18**, 285-298, (2017).

62. Martin, E. W. *et al.* Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* **367**, 694-699, (2020).

63. Tuteja, N. *et al.* Unraveling DNA helicases. Motif, structure, mechanism and function. *Eur J Biochem* **271**, 1849-1863, (2004).

64. Tuteja, N. *et al.* DNA helicases: the long unwinding road. *Nat Genet* **13**, 11-12, (1996).

65. Bochman, M. L. *et al.* DNA secondary structures: stability and function of G-quadruplex structures. *Nat Rev Genet* **13**, 770-780, (2012).

66. Luong, T. T. *et al.* Role and Regulation of the RECQL4 Family during Genomic Integrity Maintenance. *Genes (Basel)* **12**, (2021).

67. Gao, C. *et al.* Dishevelled: The hub of Wnt signaling. *Cell Signal* **22**, 717-727, (2010).

68. Lee, H. J. *et al.* Conformational change of Dishevelled plays a key regulatory role in the Wnt signaling pathways. *Elife* **4**, e08142, (2015).

69. Gammons, M. V. *et al.* Wnt Signalosome Assembly by DEP Domain Swapping of Dishevelled. *Mol Cell* **64**, 92-104, (2016).

70. Qian, Y. *et al.* Structural basis of Frizzled 4 in recognition of Dishevelled 2 unveils mechanism of WNT signaling activation. *Nat Commun* **15**, 7644, (2024).

Structure, Vol. 13, 1849–1858, December, 2005, ©2005 Elsevier Ltd All rights reserved. DOI 10.1016/j.str.2005.08.014

The Structure of the Human ERCC1/XPF Interaction Domains Reveals a Complementary Role for the Two Proteins in Nucleotide Excision Repair

Konstantinos Tripsianes,[1] Gert Folkers,[1] Eiso AB,[1] Devashish Das,[1] Hanny Odijk,[2] Nicolaas G.J. Jaspers,[2] Jan H.J. Hoeijmakers,[2] Robert Kaptein,[1] and Rolf Boelens[1,*]

[1]Department of NMR Spectroscopy
Bijvoet Center for Biomolecular Research
Utrecht University
Padualaan 8
3584 CH Utrecht
The Netherlands

[2]Department of Cell Biology and Genetics
Erasmus Medical Center
3000 DR Rotterdam
The Netherlands

Summary

The human ERCC1/XPF complex is a structure-specific endonuclease with defined polarity that participates in multiple DNA repair pathways. We report the heterodimeric structure of the C-terminal domains of both proteins responsible for ERCC1/XPF complex formation. Both domains exhibit the double helix-hairpin-helix motif $(HhH)_2$, and they are related by a pseudo-2-fold symmetry axis. In the XPF domain, the hairpin of the second motif is replaced by a short turn. The ERCC1 domain folds properly only in the presence of the XPF domain, which implies a role for XPF as a scaffold for the folding of ERCC1. The intersubunit interactions are largely hydrophobic in nature. NMR titration data show that only the ERCC1 domain of the ERCC1/XPF complex is involved in DNA binding. On the basis of these findings, we propose a model for the targeting of XPF nuclease via ERCC1-mediated interactions in the context of nucleotide excision repair.

Introduction

Genomes are vulnerable to a plethora of DNA-damaging agents, of both endogenous and environmental origin, which can compromise vital processes such as transcription and replication. To cope with these threats, all organisms have developed a complex defense network, including a variety of repair mechanisms, each suited for specific classes of DNA lesions (for a review, see Hoeijmakers, 2001). Any defect in this network invariably leads to genomic or chromosomal instability, either spontaneous or after genotoxic challenge. On the cellular level, unrepaired damage is considered the driving force of carcinogenesis as well as organismal ageing.

A remarkable structure-specific DNA nuclease family in eukaryotes is represented by the heterodimeric ERCC1/XPF complex. This enzyme is known to specifically cleave near junctions between single-stranded (ss) and duplex (ds) DNA (de Laat et al., 1998a), where the single strand has a 5′-3′ polarity by recognizing the DNA tertiary structure itself and not the nucleotide sequence. The structure-specific endonuclease ERCC1/XPF performs an essential late step in the nucleotide excision repair (NER) process (Volker et al., 2001), where it nicks the damaged DNA strand at the 5′ side of a helix-distorting lesion. This action of ERCC1/XPF, together with XPG, a complementary structure-specific endonuclease attacking on the other side of the lesion at the ss-ds transition, is required for removal of the damaged segment and the subsequent resynthesis of the segment-spanning gap. A significant contribution of the ERCC1 subunit to NER is interaction with the XPA protein, which binds to the lesion in conjunction with the strand-opening helicase complex TFIIH and the single-strand binding protein RPA (Araujo and Wood, 1999; de Laat et al., 1999; Riedl et al., 2003). As such, it directs its XPF partner to a site of NER action and contributes to the correct positioning of the strand incisions (Li et al., 1995).

Inherited defects in the NER process cause the serious prototype repair disorders xeroderma pigmentosum (XP) and Cockayne syndrome (CS), highlighting an extreme risk of UV-induced skin cancer and many features of accelerated ageing, respectively (de Boer and Hoeijmakers, 2000). Mice targeted for knockout (KO) versions of genes involved in NER usually reflect the human pathology of XP and CS to a considerable extent. In contrast, ERCC1- or XPF-targeted KO mice display a much more severe phenotype with multiorgan involvement, severe runting, and death before weaning (Tian et al., 2004; Weeda et al., 1997). This observation points to other functions of this endonuclease outside the context of NER. In line with this finding, XPF mutations associated with XP are relatively rare and always hypomorphic, whereas a first case of human ERCC1 deficiency remains to be reported.

One of these additional roles of ERCC1/XPF is in the recombinational repair of interstrand crosslinks, where it is required for a yet unidentified late step (Niedernhofer et al., 2004). ERCC1- or XPF-deficient hamster mutants are therefore exquisitely hypersensitive to DNA crosslinking agents (Prasher et al., 2005), much more than they are to UV-induced pyrimidine dimers, the classical substrates for NER. This function in crosslink repair may also explain the observations that expression levels of ERCC1/XPF in non-small cell lung cancer may anticorrelate with the response to antitumor therapy with the DNA crosslinker *cis*-platinum (Simon et al., 2005).

In addition, the homologous complexes of *S. cerevisiae* (RAD10/RAD1) and *Drosophila* (ERCC1/MEI9) are known to function in both mitotic and meiotic recombination. Indeed, ERCC1/XPF is also absolutely required for targeted recombination in mouse embryonic stem cells (Niedernhofer et al., 2001). Finally, there is recent evidence that ERCC1/XPF mediates genome integrity by yet another way, since it can prevent uncapped telomeres from creating chromosomal end-to-end fusions or double-minute chromosomes (Zhu et al., 2003).

For the action of ERCC1/XPF, heterodimer formation is essential, where the actual nuclease domain is

*Correspondence: r.boelens@chem.uu.nl

contributed by the XPF subunit (Enzlin and Scharer, 2002) and the presence of ERCC1 is indispensable for nuclease activity. Details of ERCC1's function in the actual nucleolytic action are still unknown. The XPF homolog of archaebacteria acts as a homodimer (Newman et al., 2005; Nishino et al., 2003) in the absence of an ERCC1 homolog. However, in eukaryotic cells, formation of the heterodimeric complex is required for the stability of both components. The interaction of the subunits depends on the presence of their C-terminal regions (de Laat et al., 1998b), which both carry putative double helix-hairpin-helix $(HhH)_2$ motifs (Figure 4A). These domains are known to mediate DNA binding, and it has been suggested that they are responsible for the correct positioning of the endonuclease at junctions between ds- and ss-DNA where the single strand moves 5′ to 3′ away from the junction (de Laat et al., 1998a). The phenomenon of mutual dependence has hampered overproduction and study of the separate human subunits in bacterial and insect expression systems in the past. Moderate amounts have been generated in baculovirus-infected insect cells, provided both subunits are expressed simultaneously (Enzlin and Scharer, 2002).

Since the affinity of ERCC1/XPF for single-strand to double-strand transitions is the common denominator in all of its known functions, we have set out to isolate and characterize the heterodimeric ERCC1/XPF interface containing the $(HhH)_2$ motifs.

In this study, we report the three-dimensional structure of human ERCC1/XPF consisting of the last 83 residues of XPF and the last 78 residues of ERCC1, which are sufficient for heterodimerization. Analysis of the structure allows us to map in atomic detail the interacting residues and explain the role of the C-terminal region of the ERCC1 $(HhH)_2$ domain in proper complex formation (de Laat et al., 1998b). NMR titrations established the HhH motifs of ERCC1 as the DNA binding unit, an observation that revises our understanding of the tasks each subunit performs.

Results

Purification and Biophysical Characterization of the C-Terminal Parts of XPF and ERCC1

Initial attempts to overexpress the C-terminal part of ERCC1 in *E. coli* cells resulted in insoluble and aggregated protein. Attempts to resolubilize the protein domain in various manners failed. This might be explained by our previous findings indicating that ERCC1 can only form a functional fold in complex with XPF. Therefore, the C-terminal parts of XPF and ERCC1 were coexpressed by using a bicistronic expression vector encoding the two subunits, with XPF being the first in order and ERCC1 containing a His_6 tag at the C terminus. We obtained soluble proteins with very high expression yield, and the elution profile in the gel filtration resulted in one species with the predicted size of the heterodimeric complex. Furthermore, the protein fingerprint spectrum (^{15}N-^{1}H HSQC) revealed the expected number of residues, confirming complex formation and a compact fold. These data are in accordance with the majority of the biochemical observations that illustrate the significance of the close partnership for the smooth functioning of ERCC1 and XPF proteins (Sijbers et al., 1996a).

A vast amount of free XPF was present in the unbound fraction after Ni^{2+} chromatography. We also purified the free XPF fraction and recorded its ^{15}N-^{1}H HSQC, which showed a large dispersion, indicating a folded conformation (data not shown). Urea unfolding and refolding of the complex of the ERCC1 and XPF domains results in free XPF, as can be concluded from the ^{15}N-^{1}H HSQC spectrum, and a precipitate of the ERCC1 domain. Refolding into a heterodimeric ERCC1/XPF complex was possible when His_6-tagged ERCC1 was immobilized on a Ni^{2+} affinity column and XPF was circulated in excess over this column while gradually decreasing the urea concentration (Figure 1). From that we conclude that the ERCC1 domain can only fold in the presence of the XPF domain under our in vitro conditions.

Structure Determination

The solution structure of the 19 kDa complex between the C-terminal domain of ERCC1 (residues 220–297) and the C-terminal domain of XPF (residues 823–905) was solved by heteronuclear double- and triple-resonance NMR spectroscopy by using uniformly ^{15}N- and $^{15}N/^{13}C$-labeled protein. The structure was determined on the basis of 4547 experimental NMR restraints and 197 dihedral angle restraints. A large number of intermolecular distance restraints was collected (442), which permitted the positioning of the two subunits with respect to each other. No distance violations larger than 0.4 Å or dihedral angle violations larger than 5° were found. A summary of the structural and restraint statistics is given in Table 1. Figure 2A shows an overlay of the 20 lowest-energy conformers obtained after the structure calculations.

Structure Description

The structure of the core of the complex, comprising residues 234–294 of ERCC1 and residues 836–895 of XPF, exhibits a global pseudosymmetry with a very similar architecture of the associated partners. The fold of both subunits is built up by helical secondary elements. In particular, the fold of ERCC1 is a characteristic example of the double helix-hairpin-helix $(HhH)_2$ motif (Shao and Grishin, 2000), which is present in a variety of protein families involved in non-sequence-specific DNA binding such as DNA-polymerases, ligases, and nucleases. Each motif forms into a pair of antiparallel α helices connected by a hairpin-like loop. The ERCC1 helices α and β form the first HhH element, and helices δ and ε form the second one connected by a short helix, γ. The angle between the helical vectors on the first and second pair is 68 and 49°, respectively. On the other side, the XPF fold is almost the same, except for a minor, but significant, difference. Although the relative orientation of the XPF helices agrees well with the $(HhH)_2$ fold, only the first hairpin is present. It is formed by helices α' and β' with an angle of 51° (helices of XPF are indicated by a prime sign). The hairpin between helices δ' and ε' has been replaced by a β turn of only three residues. These helices are inclined at an angle of 55°. Although XPF retains only one complete HhH element, the global fold still resembles the $(HhH)_2$ structure (Figures 2B and 2C). The rmsd of the individual ERCC1

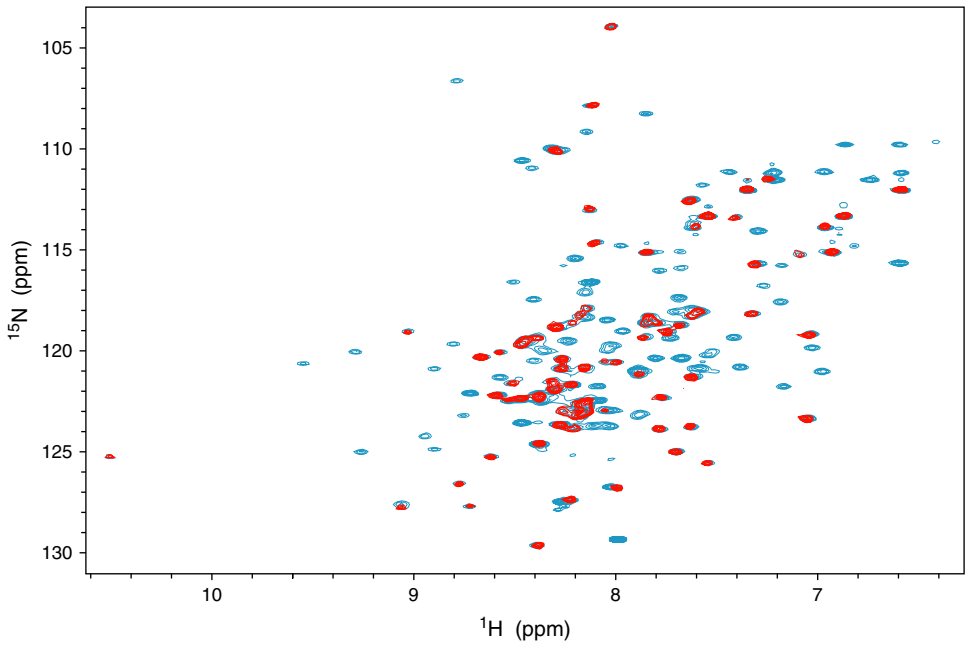


Figure 1. Refolding of the ERCC1/XPF Complex

The figure shows the ^{15}N-^{1}H HSQC spectrum of the ERCC1/XPF complex (blue) and the ^{15}N-^{1}H HSQC spectrum of ^{15}N-labeled ERCC1 (red) refolded in the presence of unlabeled XPF. The overlay of the crosspeaks of ERCC1 in the two spectra confirms that the refolded complex forms the native conformation (for more details, see Experimental Procedures).

and XPF structures is 2.5 Å (65 Cα atoms), with a Z-score of 8.1.

Protein-Protein Interactions

The two domains of ERCC1 and XPF have a large interaction surface with an area of 1534 Å^2. The orientation of helices α and γ from both subunits resembles a four-helical bundle, although the γ helices are very short. Only helix γ′ of XPF deviates from the parallel orientation, and it is tilted by 35° with respect to its ERCC1 mate (Figure 3B). Both ERCC1 and XPF helical pairs form relatively flat and extremely hydrophobic surfaces that are shielded in the protein-protein interface. This central hydrophobic core seems to be the main contributor of the association, since it confers more than half of the interaction surface area. A key determinant of the nonbonded interactions is Phe840 of XPF, which protrudes in the middle of the symmetric helical array (Figures 3A and 3B). The aromatic ring of Phe840 makes contacts with Cys238, Thr241, Leu260, and Leu289 of ERCC1, which are located in helices α, γ, and ε. The rest of the intermolecular contacts are restricted to residues that belong to symmetry-related helices. Lys843 in helix α′ of XPF interacts via its aliphatic side chain with Glu237, Cys238, and Thr241 of ERCC1 helix α. In half of the structures in the ensemble, the carboxyl of Lys843 of XPF forms bifurcated hydrogen bonds to the guanidinium group of ERCC1 Arg234. Moreover, the side chain amide group of the same Lys forms a salt bridge with the carboxyl side chain of Glu237 of ERCC1. Regarding the short γ helices, hydrophobic interactions are present between Ile264 of ERCC1 and Ala863 and Ala866 of XPF. The hydrophobic core additionally contains Phe889 of XPF and Val288 and Leu289 of ERCC1.

In previous studies, it has been noted that the C-terminal residues of each subunit were essential for complex formation (de Laat et al., 1998b). The structure clarifies the important role of the two Phe residues. The ring of ERCC1 Phe293 fits perfectly into a hydrophobic cavity that is formed by the HhH motif of XPF and consists of residues Pro837, Gln838, Leu841, Met856, Asn861, and Ile862. The hydrophobic pocket has a large contact surface of 280 Å^2. In this region, the backbone of ERCC1 forms three hydrogen bonds with XPF that anchor the side chain of Phe293 to the XPF cavity. The side chains of Tyr833 from XPF and Leu294 from ERCC1 cover the entry of the cavity and lock the pocket, with Phe293 buried in it (Figure 3C). A similar pocket is built up by the first HhH motif of ERCC1, which accommodates, in this case, the ring of XPF Phe894. The pocket consists of ERCC1 residues Phe231, Val232, Val235, and Leu254, and the contact surface is smaller than that of XPF and has an area of 220 Å^2. It is stabilized by five intermolecular hydrogen bonds from the XPF backbone (Figure 3D). Residues that form the interaction surface are conserved in all mammals, while generally conservative substitutions are found in other eukaryotes, thereby preserving the hydrophobic nature of the complex.

Helix-Hairpin-Helix Motifs and DNA Binding

A DALI search in the Protein Data Bank using the individual protomers as queries detected the same set of homologous structures for both ERCC1 and XPF. In all cases, ERCC1 showed better scores, because it exhibits the canonical $(HhH)_2$ motif. The $(HhH)_2$ domain of ERCC1 closely resembles the $(HhH)_2$ domain of the archaeal homodimeric XPF from *Aeropyrum pernix* (PDB code 2BGW; rmsd 1.9 Å) (Newman et al., 2005) and the

Table 1. Structural Statistics of the Structure Ensemble of the Heterodimeric ERCC1/XPF Protein

Rmsd (Å) with respect to mean (backbone/heavy)	
ERCC1	0.28 ± 0.06/0.80 ± 0.11
XPF	0.30 ± 0.07/0.78 ± 0.13
Complex	0.33 ± 0.06/0.81 ± 0.10
Number of experimental restraints	
Intraresidue NOEs	876
Sequential NOEs ($\|i - j\| = 1$)	1210
Medium-range NOEs ($1 < \|i - j\| < 4$)	1252
Long-range NOEs ($\|i - j\| > 4$)	767
Interprotein	442
Total NOEs	4547
Dihedral angle restrains	197
Restraint violations	
NOE distances with violations >0.4 Å	0.00 ± 0.00
Dihedrals with violations >5°	0.00 ± 0.00
Rmsd for experimental restraints	
All distance restraints (4547) (Å)	0.0186 ± 0.0008
Torsion angles (197) (°)	0.5555 ± 0.049
CNS energies after water refinement	
E_{vdw} (kcal/mol)	−604 ± 15
E_{elec} (kcal/mol)	−6482 ± 101
Rmsd (Å) from idealized covalent geometry	
Bonds (Å)	0.01 ± 0.00
Angles (°)	1.29 ± 0.03
Impropers (°)	1.39 ± 0.06
Ramachandran analysis	
Residues in most favored regions (%)	94.37 ± 0.93
Residues in additionally allowed regions (%)	5.59 ± 0.97
Residues in generously allowed regions (%)	0.04 ± 0.20
Residues in disallowed regions (%)	0.00 ± 0.00

Residues 227–294 of ERCC1 and residues 831–896 of XPF were used.

holiday junction-recognizing $(HhH)_2$ domain of RuvA (PDB code 1C7Y; rmsd 1.9 Å) (Ariyoshi et al., 2000).We also noticed significant similarity to the C-terminal domain of the bacterial repair protein UvrC (PDB code 1KFT; rmsd 2.1 Å) (Singh et al., 2002), which exhibits the same double HhH organization (Figure 4B). We have shown before that the UvrC $(HhH)_2$ domain is required and sufficient for recognition of the single-to-double strand junction. To address whether the ERCC1/XPF HhH domains mediate DNA binding, we performed electrophoretic mobility shift assays.

Weak binding (K_d of 39 ± 13 μM) was detected on a stem-loop DNA substrate that was shown to be the optimal substrate for ERCC1/XPF in nuclease activity assays (de Laat et al., 1998a). We found that binding is specific for probes that contain single-double strand junctions, as complexes were also formed with a bubble and fork substrate, whereas no binding was detected on dsDNA or ssDNA probes (Figure S1; see the Supplemental Data available with this article online). Though weaker, the binding preference of the isolated $(HhH)_2$ interacting domains is similar to the observed in vitro incision preference of full-length recombinant ERCC1/XPF protein (de Laat et al., 1998a), which shows a K_d of 100 nM for the same stem-loop substrate (unpublished data). In this respect, it is likely that other domains of the native complex contribute to DNA binding as well. Furthermore, it has been shown before that recruitment of ERCC1/XPF in the context of NER is stabilized by various other proteins, including XPA-RPA proteins and the presence of TFIIH and XPG in the preincision complex (Riedl et al., 2003; Volker et al., 2001).

To determine the interaction surface for DNA binding, ^{1}H-^{15}N HSQC spectra of the protein were recorded with successive additions of a stem-loop DNA substrate. Although the chemical shift perturbations that we observed upon titration are relatively small, the most prominent effects are clustered in the second hairpin of ERCC1, with a few in its first hairpin (Figure 4C and Figure S2). In particular, the maximum amide chemical shift perturbations were detected in Gly276 and Gly278 along with the intervening Leu277 forming the classical GhG hairpin. In the crystal structures of protein-DNA complexes containing these hairpin motifs, the glycines form hydrogen bonds with adjacent phosphate oxygens of the DNA, and a positive residue after the second Gly makes polar contacts with the phosphate backbone. Lys281 of ERCC1 is also shifted significantly, suggesting the involvement of this residue in a protein-DNA contact. The first hairpin of ERCC1 is more distinct in terms of sequence composition and lacks the GhG signature. Unfortunately, due to overlap, we were not able to follow the shifts of three critical residues of the first hairpin, Ser244, Lys247, and Thr248. However, the shifts at Val245 and Asn246 demonstrate that this hairpin is also in contact with DNA. There are also randomly distributed shifts both on the ERCC1 and the XPF side. They probably reflect indirect perturbation effects possibly involving small rearrangements in the hydrophobic packing induced by DNA binding. No shifts, however, are observed in the hairpin of XPF, strongly suggesting that this subunit does not directly participate in DNA recognition under the present in vitro conditions. Furthermore, the positive charge mainly in the hairpins of ERCC1 agrees well with the ERCC1 residues found in the proximity of DNA in our experiments (Figures 4D and 4E).

Discussion

ERCC1 Forms an Obligate Heterodimer with XPF

In this study, we show that the driving force of complex formation is mainly hydrophobic interaction, and we show that this yields a compact architecture that enhances the stability of each subunit. Both domains adopt similar structures in the final complex related by a pseudo-2-fold symmetry axis, with ERCC1 displaying the canonical $(HhH)_2$ fold and XPF displaying a very similar fold. The DNA binding domain from ERCC1 resembles the archaeal XPF more closely than the eukaryotic XPF does. The opposite is true for the flanking endonuclease domain, which is inactivated in ERCC1 protein (Figure 4A). In archaebacteria, ERCC1 is absent, while in all sequenced eukaryotes, ranging from fungi, parasites, plants, and invertebrates to vertebrates, ERCC1 and XPF homologs have been found. Furthermore, the archaeal orthologous repair protein, like the human protein, forms a dimeric structure. The structural resemblances of human ERCC1 and archaeal XPF with regard to the $(HhH)_2$ domains agree with the suggested common origin for the eukaryotic ERCC1 and XPF proteins (Gaillard and Wood, 2001).

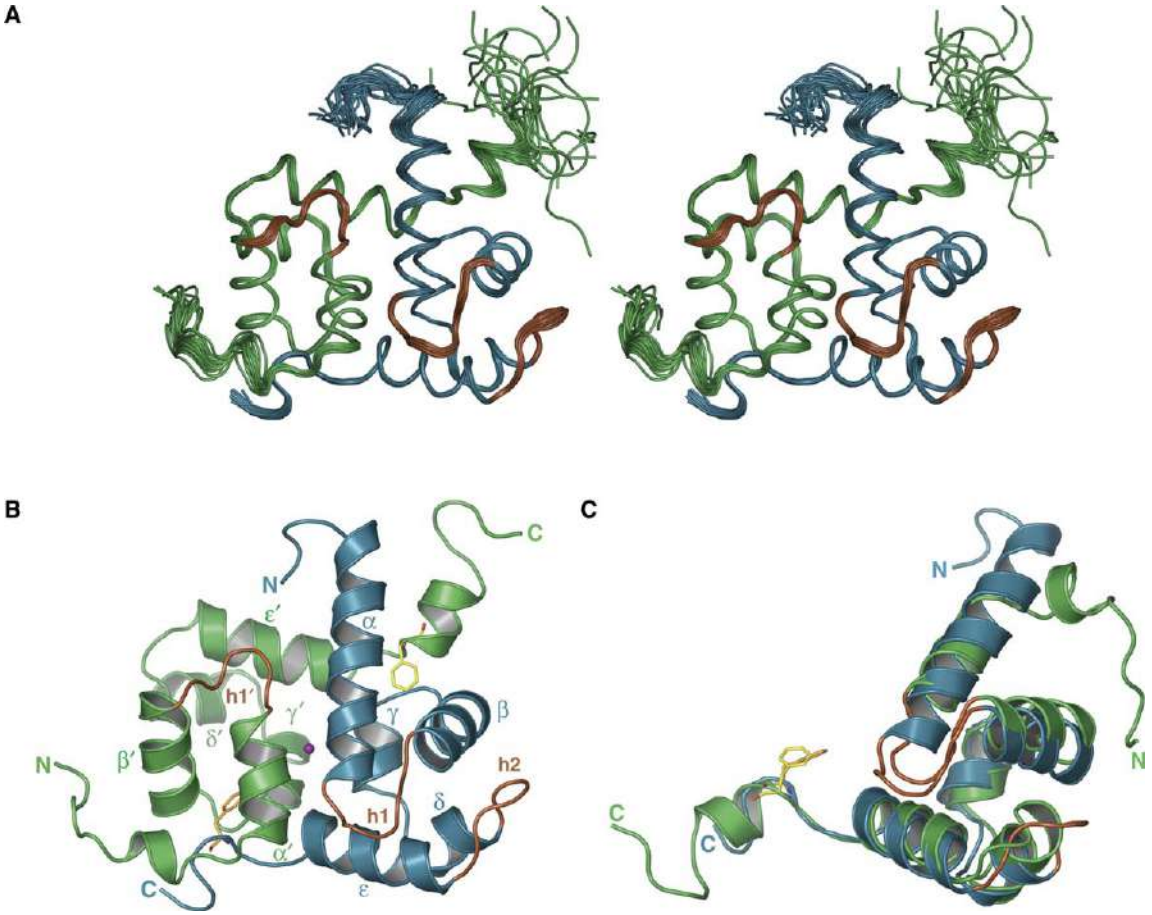


Figure 2. Three-Dimensional Structure of the C-Terminal Domains of the Human ERCC1/XPF Complex

(A) Backbone stereoview of the ensemble of the final 20 structural conformers. ERCC1 is colored blue, XPF is colored green, and the hairpins in both subunits are brown.

(B) Cartoon representation of the lowest-energy model. The purple sphere depicts the center of the pseudo-2-fold symmetry axis. Helices are denoted with Greek letters, from α to ε for ERCC1 and from α' to ε' for XPF, while hairpins are indicated as h1, h2, and h1′. Also shown are Phe residues at the C termini of each subunit. Phe293 of ERCC1 is colored orange, and Phe894 of XPF is colored yellow.

(C) ERCC1 and XPF are superimposed, showing the overall fold similarity. Color conventions are as in (B).

The requirement for the formation of a stable ERCC1 through heterodimer association reflects the importance of the biological function of ERCC1. We have demonstrated that the double hairpin motif on the ERCC1 side mediates DNA binding to a stem loop substrate, an equivalent of the NER bubble, which is the most preferred substrate of eukaryotic ERCC1/XPF. Although in vivo the folding process is probably more complicated and could involve chaperone-mediated posttranslational interactions with XPF, our in vitro biochemical data support the finding that ERCC1 functionality is strictly dependent on the XPF domain as a folding scaffold. While this manuscript was in preparation, a biophysical characterization of the same interacting domains was reported (Choi et al., 2005). Although in this study the ERCC1 and XPF C-terminal domains were expressed separately, stabilization of ERCC1 was achieved only in the presence of XPF, in full accordance with our refolding scheme.

This finding, together with the high-resolution structure presented here, is related to two studies that have reported sensitivity of the native complex to short C-terminal ERCC1 truncations and highlighted a role for Phe293 in stability and function (de Laat et al., 1998b; Sijbers et al., 1996b). The structure of the complex suggests that locking of this residue to its partner pocket links directly to the formation of the second hairpin motif of ERCC1 and indirectly to DNA recognition capability and function. Evolutionary conservation of Phe293 in all ERCC1 family members underscores its importance. Deletion of the Phe "hook" leads to loss of functionality, presumably because a large interacting area is abolished (Sijbers et al., 1996b).

DNA Binding Is Mediated by the HhH Domain of ERCC1

The observed small chemical shift changes in the core of the ERCC1/XPF complex upon addition of DNA (Figure 4) suggest that small, local conformational changes are required to accommodate the DNA substrate for the human ERCC1/XPF heterodimer. In this respect, we note that the $(HhH)_2$ motif of ERCC1 resembles more closely the corresponding domain of the archaeal homodimeric XPF bound to DNA than that of the free archaeal XPF structure (Newman et al., 2005). It is known that HhH domains recognize various substrates, but

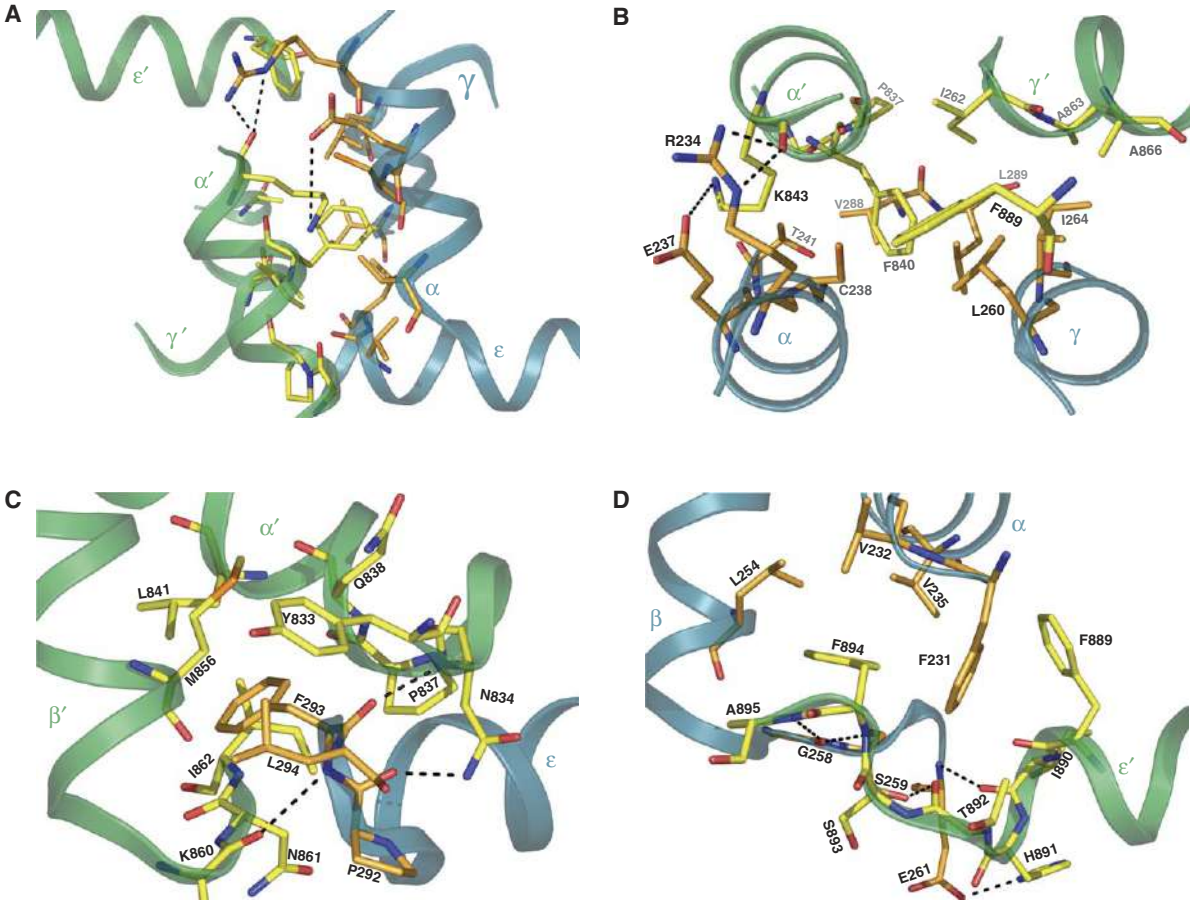


Figure 3. Protein-Protein Interactions along the C-Terminal Domains of ERCC1 and XPF

(A) View of the central interacting core. Helices α in the front, helices γ at the back, and the tips of helices ε that run perpendicularly build up the interacting surface. Residues participating in contacts are depicted in stick representation. The backbone of ERCC1 is shown in blue, XPF is shown in green, and interacting amino acids from either side are shown in orange and yellow, respectively. Intermolecular hydrogen bonds are indicated with black dotted lines.

(B) Top view of the central core. For the sake of clarity, helices ε have been removed.

(C and D) (C) The tight hydrophobic packing of ERCC1 Phe293 to its XPF cavity and (D) the same for Phe894 of XPF in the hydrophobic pocket of ERCC1. In (B), (C), and (D), amino acids are numbered according to the native sequences.

what really determines the specificity remains largely unclear. The crystal structures of the archaeal XPF homodimer and bacterial RuvA bound to DNA have revealed a similar pattern of contacts between GhG hairpins and the minor groove of DNA. Also, in bacterial UvrC (Singh et al., 2002), similar or identical residues that form the double HhH motif were found to be crucial for DNA binding. We observed the same residue contacts for the second hairpin of ERCC1 in our chemical shift perturbation experiments. Even on the first hairpin, which is less homologous with the classic GhG hairpins, the perturbed residues agree very well with the crystallographic complexes. Sequence analysis of ERCC1 and XPF members revealed that residues critically involved in DNA binding in our NMR data are strictly conserved in the ERCC1 family, but are mostly absent in XPF. Remarkably, most of these ERCC1 residues are present and conserved in archaeal XPF and are shown to be in contact with DNA (Newman et al., 2005).

Our observation that only the $(HhH)_2$ domain of ERCC1 binds to DNA points to an interesting difference in topology between human ERCC1/XPF and the archaeal homodimeric XPF proteins. The crystal structure of the archaeal XPF suggests that $(HhH)_2$ and the nuclease domain of the same protomer are involved in DNA contacts. In the human case, it has been shown that only XPF carries a nuclease domain preceding its C-terminal domain (Enzlin and Scharer, 2002). Thus, for ERCC1/XPF, it appears that DNA binding is located in one subunit (ERCC1) and that the nuclease activity is located in the other subunit (XPF).

Functional Implications of the Intimate Association between the HhH Domains of ERCC1 and XPF

ERCC1/XPF is the last factor arriving at the DNA damage site, resulting in the mature preincision NER complex. It has been shown that the binding of ERCC1/XPF depends on a multitude of sequential prerequisites, including the TFIIH DNA unwinding step (Mone et al., 2004), the presence of XPG at the opposite cleavage site, and the binding of XPA-RPA (Riedl et al., 2003; Volker et al., 2001). Indeed, the middle domain of ERCC1 can interact specifically with XPA (Bessho et al., 1997; Li et al., 1995). Furthermore, we have shown here that the ERCC1

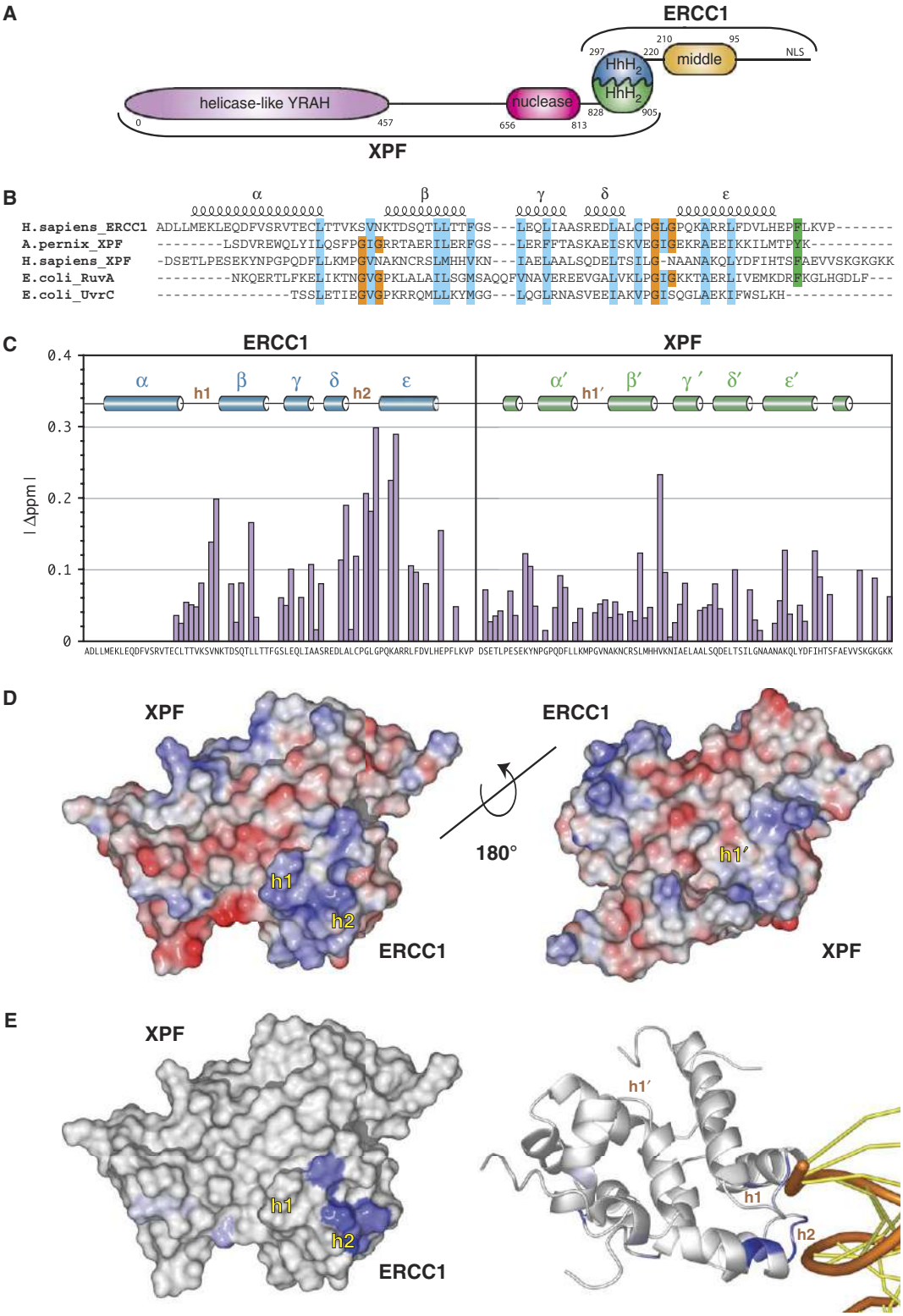


Figure 4. DNA Binding by the ERCC1/XPF Complex

(A) Domain organization of the ERCC1/XPF complex.

(B) Sequence alignment of selected $(HhH)_2$ domains from the three kingdoms. Hydrophobic residues are highlighted in cyan, Gly that belong to hairpin motifs are highlighted in orange, and the aromatic residues at the end of the domains are highlighted in green. The observed secondary elements of ERCC1 are indicated above the sequences.

(C) Normalized chemical shift changes upon DNA titration versus the ERCC1 and XPF sequences. Missing bars indicate either Pro residues or unresolved chemical shifts due to peak overlap.

$(HhH)_2$ domain is sufficient to mediate DNA binding (Figure 4). We propose that ERCC1 combines DNA binding activity and specific interactions with XPA to recruit the endonuclease activity of XPF to the incision complex. This guarantees that XPF cleavage is restricted only to correctly processed DNA templates. This model for ERCC1/XPF is in full accordance with the assembly of NER factors at sites of DNA damage in a sequential and interdependent manner. Our interpretation emphasizes the functional distinction of the tightly associated partners with ERCC1 making all the contacts with DNA and XPA that will target the nuclease of XPF to perform the strand cleavage 5′ upstream of the lesion.

Experimental Procedures

Protein Expression and Purification

Earlier experiments revealed that ERCC1 was expressed in inclusion bodies and that resolubilization of ERCC1 with urea and refolding did not result in a properly folded protein. In addition, we have mapped the interaction domains between ERCC1 and XPF to residues 224–297 and 814–905 for ERCC1 and XPF, respectively (de Laat et al., 1998b). To obtain a stable heterodimeric complex with minimal flanking sequences, putative domain boundaries were determined by using bioinformatics tools as described previously (Folkers et al., 2004). For both domains, several N-terminal deletion constructs around the previously identified domain (ERCC1: 220, 224, 234; XPF: 813, 823, 832) were tested by using in vitro interaction assays (described in de Laat et al., 1998b). From these interaction experiments, we concluded that the minimal interaction domain is composed of residues 220–297 and 823–905 for ERCC1 and XPF, respectively. Therefore, we constructed a bicistronic expression vector in which we first cloned the XPF sequence 823–905 (all oligonucleotide sequences used are available upon request) in the NcoI-BamH1 site of pET28b. Subsequently, in the BamH1-XhoI site of this XPF expression construct, the ERCC1 fragment 220–297, extended with a 27 base pair linker containing an optimal internal ribosome binding site, was cloned, resulting in a bicistronic expression construct in which only the ERCC1 domain contains the His_6 tag.

Overexpression and His tag purification of the ^{15}N and $^{15}N/^{13}C$ ERCC1/XPF complexes was performed as described before (Folkers et al., 2004), except that elution was performed with 400 mM EDTA. After elution, dialysis against NMR buffer (50 mM $NaPO_4$, 100 mM NaCl [pH 7.0]) was performed by using a 3K cutoff dialysis tube (SpectraPor). Subsequently, the complex was loaded on a Superdex 75 column in the same buffer. The ERCC1/XPF complex eluted from this column as a single species at an elution time as expected for a heterodimer of this size. Even when large amounts of the complex were loaded on the column, no monomeric ERCC1 or XPF was detected, and, together with the observed symmetric elution profile for the ERCC1/XPF complex, we conclude that ERCC1 and XPF are in tight association. After concentration of this fraction to 0.5–1.5 mM with ultrafiltration with a 3K filter (amicon; Milipore), protease inhibitor cocktail (Complete EDTA free; Roche) was added.

For on-column refolding experiments, the ERCC1/XPF complex (^{15}N labeled) was first loaded on the nickel-charged metal chelating column of 7.8 ml (Poros MC20; Applied Biosystems) under native conditions (50 mM $NaPO_4$, 150 mM NaCl, 20 mM imidazole, 1 mM β-mercaptoethanol, 0.2 mM PMSF [pH 8.0]) by using BioCad Vision (Applied Biosystems). Subsequently, the XPF moiety was eluted from the column by using a denaturing buffer (native buffer containing 6 M urea). The 20 ml of eluted XPF fraction (from a nonlabeled elution) was reloaded with a flow rate of 5 ml/min by continuous recycling of this XPF-containing fraction on the column while simultaneously mixing the eluate with an increasing amount of native buffer. In this procedure, the urea was gradually decreased from 6 M to a final 200 mM concentration over a total volume of 600 ml. After washing with native buffer, the ERCC1/XPF complex was eluted with 400 mM of EDTA and was purified as described above.

DNA Binding

Electrophoretic mobility shift assays were performed in a binding buffer containing 50 mM $NaPO_4$, 100 mM NaCl, 20 mg/ml BSA, and 10% (v/v) glycerol, by using radio-labeled stem-loop, bubble, fork, dsDNA (10 bp), and ssDNA (20 mer) probes, as described before (Singh et al., 2002). Unless otherwise indicated, for all experiments, 50 μM ERCC1/XPF was used. For supershift, 0.5 ml HIS-probeHRP conjugate (Pierce) was added to the reaction mixture prior to the addition of probe. MagneHis (Promega) was added to the reaction mixture and cleared by using a magnetic stand, either prior to the addition of DNA or after complex formation, yielding identical results.

For DNA titrations, we used 0.125 mM of ^{15}N-labeled ERCC1/XPF complex in NMR buffer and added increasing amounts of a hairpin sequence in an identical buffer, with 22 unpaired bases (5′-GCCAGC GCTCGGTTTTTTTTTTTTTTTTTTTTTTCCGAGCGCTGGC-3′) until a final concentration of 0.25 mM was reached. ^{15}N-^{1}H HSQC spectra were recorded for the different titration points. Normalized chemical shift changes (δ) for the nonoverlapping residues were calculated by using the equation: $\delta = ([5\delta_{HN}]^2 + [\delta_N]^2)^{0.5}$ (Grzesiek et al., 1996).

NMR Spectroscopy

All NMR spectra were recorded on Bruker AVANCE 700 and AVANCE 900 spectrometers equipped with triple-resonance gradient probes at 295.5 K, using [U-^{13}C, ^{15}N]- and [U-^{15}N]-labeled proteins. Backbone and side chain ^{1}H, ^{15}N, and ^{13}C resonances for ERCC1/XPF were assigned by using the following set of 3D triple-resonance experiments: HNCACB, CBCA(CO)NH, HN(CA)HA, HBHA(CBCACO)NH, HNCO, HN(CA)CO, C(CO)NH-TOCSY, H(CCO)NH-TOCSY, H(C)CH-TOCSY, and (H)CCH-TOCSY spectra acquired at 700 MHz. A full set of 3D heteronuclear-edited NOE spectra were then recorded at 900 MHz for structure determination: 3D NOESY-(^{13}C,^{1}H)-HSQC (τ_{mix} = 75 ms; tuned once for aliphatic and once for aromatic ^{13}C), 3D NOESY-(^{15}N,^{1}H)-HSQC (τ_{mix} = 80 ms), 3D (^{13}C)-HMQC-NOESY-(^{13}C,^{1}H)-HSQC (τ_{mix} = 75 ms), and 3D (^{13}C)-HMQC-NOESY-(^{15}N-^{1}H)-HSQC (τ_{mix} = 80 ms, tuned for aliphatic ^{13}C). These were complemented by homonuclear 2D NOE spectra recorded without (τ_{mix} = 80 ms) and with H[^{15}N] suppression in F2 (τ_{mix} = 80 ms). The triple-resonance and heteronuclear 3D experiments were performed essentially as described by (Cavanagh et al., 1996). All spectra were processed by using the NMRPipe software package (Delaglio et al., 1995) and were analyzed with Sparky (Goddard and Kneller, 2004).

Structure Calculations

Automatic NOE assignment and structure calculations were performed by using the CANDID module of the program CYANA (Güntert et al., 1997; Herrmann et al., 2002). The quality of the structures was improved in an iterative procedure in which CANDID runs were followed by manual inspection of the preliminary structures to find additional resonance assignments from the NOE spectra. Hydrogen bond restraints were defined when they were consistent with the secondary shift data and expected NOE contacts. Manual NOE peak assignments were generally not held constant in the CANDID runs, but used to create accurate spectrum-specific chemical shift lists and to check the consistency of subsequent CANDID runs. The final CANDID run was performed with CYANA version 2.0, by using Ramachandran and side chain rotamer restraints for every cycle except for the last one. In the final cycle, fixed stereospecific assignments for prochiral groups were used if available. Finally, the set of NOE-based restraints determined by CANDID, together with restraints for 53 H bonds and 220 ϕ and ψ torsion angle restraints derived from TALOS (Cornilescu et al., 1999) were used in a water

(D) Two views of surface representation rotated by 180° and colored according to their electrostatic surface potential at ±8 kB T/e for positive (blue) or negative (red) charge potential by using the program GRASP (Nicholls, 1993).
(E) Observed chemical shift changes on the protein surface and model for the interaction of ERCC1/XPF with DNA by fitting ERCC1 on the archaeal $(HhH)_2$ domain. The blue parts denote shifting residues in both representations, and the intensity of the color corresponds to the absolute shift value (cut-off value of 0.1).

refinement run by using CNS (Brünger et al., 1998) according to the standard RECOORD protocol (Nederveen et al., 2005). The final structures were validated with WHATIF (Vriend, 1990) and PROCHECK (Laskowski et al., 1993; Morris et al., 1992). Molecular images were generated with PyMol/NUCCYL (Delano, 2002).

Supplemental Data

Supplemental Data including Figures S1 and S2 are available at http://www.structure.org/cgi/content/full/13/12/1849/DC1/.

Acknowledgments

The authors are grateful to Dr. Tammo Diercks for expert NMR assistance. This work was financially supported by the Research Council for the Chemical Sciences of the Netherlands Organization for Scientific Research (NWO-CW) and by the Center for Biomedical Genetics. H.O., N.G.J.J., and J.H.J.H. were supported by Euratom grants nrs FIGH-CT-2002-00207 and LSHG-CT-2005-512113.

Received: June 24, 2005
Revised: August 26, 2005
Accepted: August 28, 2005
Published: December 13, 2005

References

Araujo, S.J., and Wood, R.D. (1999). Protein complexes in nucleotide excision repair. Mutat. Res. *435*, 23–33.

Ariyoshi, M., Nishino, T., Iwasaki, H., Shinagawa, H., and Morikawa, K. (2000). Crystal structure of the holliday junction DNA in complex with a single RuvA tetramer. Proc. Natl. Acad. Sci. USA *97*, 8257–8262.

Bessho, T., Sancar, A., Thompson, L.H., and Thelen, M.P. (1997). Reconstitution of human excision nuclease with recombinant XPF-ERCC1 complex. J. Biol. Chem. *272*, 3833–3837.

Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. *54*, 905–921.

Cavanagh, J., Fairbrother, J.W., Palmer, G.A., III, and Skelton, J.N. (1996). Protein NMR Spectroscopy. (San Diego, CA: Academic Press).

Choi, Y.J., Ryu, K.S., Ko, Y.M., Chae, Y.K., Pelton, J.G., Wemmer, D.E., and Choi, B.S. (2005). Biophysical characterization of the interaction domains and mapping of the contact residues in the XPF-ERCC1 complex. J. Biol. Chem. *280*, 28644–28652.

Cornilescu, G., Delaglio, F., and Bax, A. (1999). Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J. Biomol. NMR *13*, 289–302.

de Boer, J., and Hoeijmakers, J.H. (2000). Nucleotide excision repair and human syndromes. Carcinogenesis *21*, 453–460.

de Laat, W.L., Appeldoorn, E., Jaspers, N.G., and Hoeijmakers, J.H. (1998a). DNA structural elements required for ERCC1-XPF endonuclease activity. J. Biol. Chem. *273*, 7835–7842.

de Laat, W.L., Sijbers, A.M., Odijk, H., Jaspers, N.G., and Hoeijmakers, J.H. (1998b). Mapping of interaction domains between human repair proteins ERCC1 and XPF. Nucleic Acids Res. *26*, 4146–4152.

de Laat, W.L., Jaspers, N.G., and Hoeijmakers, J.H. (1999). Molecular mechanism of nucleotide excision repair. Genes Dev. *13*, 768–785.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., and Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J. Biomol. NMR *6*, 277–293.

Delano, W. (2002). The PyMOL Molecular Graphics System. (San Carlos, CA: DeLano Scientific).

Enzlin, J.H., and Scharer, O.D. (2002). The active site of the DNA repair endonuclease XPF-ERCC1 forms a highly conserved nuclease motif. EMBO J. *21*, 2045–2053.

Folkers, G.E., van Buuren, B.N., and Kaptein, R. (2004). Expression screening, protein purification and NMR analysis of human protein domains for structural genomics. J. Struct. Funct. Genomics *5*, 119–131.

Gaillard, P.H., and Wood, R.D. (2001). Activity of individual ERCC1 and XPF subunits in DNA nucleotide excision repair. Nucleic Acids Res. *29*, 872–879.

Goddard, T.D., and Kneller, D.G. (2004). SPARKY 3 (computer program). University of California, San Francisco.

Grzesiek, S., Bax, A., Clore, G.M., Gronenborn, A.M., Hu, J.S., Kaufman, J., Palmer, I., Stahl, S.J., and Wingfield, P.T. (1996). The solution structure of HIV-1 Nef reveals an unexpected fold and permits delineation of the binding surface for the SH3 domain of Hck tyrosine protein kinase. Nat. Struct. Biol. *3*, 340–345.

Güntert, P., Mumenthaler, C., and Wüthrich, K. (1997). Torsion angle dynamics for NMR structure calculation with the new program DYANA. J. Mol. Biol. *273*, 283–298.

Herrmann, T., Güntert, P., and Wüthrich, K. (2002). Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J. Mol. Biol. *319*, 209–227.

Hoeijmakers, J.H. (2001). Genome maintenance mechanisms for preventing cancer. Nature *411*, 366–374.

Laskowski, R.A., MacArthur, M.W., Moss, D.S., and Thornton, J.M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystalogr. *26*, 283–291.

Li, L., Peterson, C.A., Lu, X., and Legerski, R.J. (1995). Mutations in XPA that prevent association with ERCC1 are defective in nucleotide excision repair. Mol. Cell. Biol. *15*, 1993–1998.

Mone, M.J., Bernas, T., Dinant, C., Goedvree, F.A., Manders, E.M., Volker, M., Houtsmuller, A.B., Hoeijmakers, J.H., Vermeulen, W., and van Driel, R. (2004). In vivo dynamics of chromatin-associated complex formation in mammalian nucleotide excision repair. Proc. Natl. Acad. Sci. USA *101*, 15933–15937.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G., and Thornton, J.M. (1992). Stereochemical quality of protein structure coordinates. Proteins *12*, 345–364.

Nederveen, A.J., Doreleijers, J.F., Vranken, W., Miller, Z., Spronk, C.A., Nabuurs, S.B., Güntert, P., Livny, M., Markley, J.L., Nilges, M., et al. (2005). RECOORD: A recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. Proteins *59*, 662–672.

Newman, M., Murray-Rust, J., Lally, J., Rudolf, J., Fadden, A., Knowles, P.P., White, M.F., and McDonald, N.Q. (2005). Structure of an XPF endonuclease with and without DNA suggests a model for substrate recognition. EMBO J. *24*, 895–905.

Nicholls, A. (1993). GRASP: Graphical Representation and Analysis of Surface Properties. (New York: Columbia University).

Niedernhofer, L.J., Essers, J., Weeda, G., Beverloo, B., de Wit, J., Muijtjens, M., Odijk, H., Hoeijmakers, J.H., and Kanaar, R. (2001). The structure-specific endonuclease Ercc1-Xpf is required for targeted gene replacement in embryonic stem cells. EMBO J. *20*, 6540–6549.

Niedernhofer, L.J., Odijk, H., Budzowska, M., van Drunen, E., Maas, A., Theil, A.F., de Wit, J., Jaspers, N.G., Beverloo, H.B., Hoeijmakers, J.H., and Kanaar, R. (2004). The structure-specific endonuclease Ercc1-Xpf is required to resolve DNA interstrand cross-link-induced double-strand breaks. Mol. Cell. Biol. *24*, 5776–5787.

Nishino, T., Komori, K., Ishino, Y., and Morikawa, K. (2003). X-ray and biochemical anatomy of an archaeal XPF/Rad1/Mus81 family nuclease: similarity between its endonuclease domain and restriction enzymes. Structure (Camb) *11*, 445–457.

Prasher, J.M., Lalai, A.S., Heijmans-Antonissen, C., Ploemacher, R.E., Hoeijmakers, J.H., Touw, I.P., and Niedernhofer, L.J. (2005). Reduced hematopoietic reserves in DNA interstrand crosslink repair-deficient Ercc1(−/−) mice. EMBO J. *24*, 861–871.

Riedl, T., Hanaoka, F., and Egly, J.M. (2003). The comings and goings of nucleotide excision repair factors on damaged DNA. EMBO J. *22*, 5293–5303.

Shao, X., and Grishin, N.V. (2000). Common fold in helix-hairpin-helix proteins. Nucleic Acids Res. *28*, 2643–2650.

Sijbers, A.M., de Laat, W.L., Ariza, R.R., Biggerstaff, M., Wei, Y.F., Moggs, J.G., Carter, K.C., Shell, B.K., Evans, E., de Jong, M.C., et al. (1996a). *Xeroderma pigmentosum* group F caused by a defect in a structure-specific DNA repair endonuclease. Cell *86*, 811–822.

Sijbers, A.M., van der Spek, P.J., Odijk, H., van den Berg, J., van Duin, M., Westerveld, A., Jaspers, N.G., Bootsma, D., and Hoeijmakers, J.H. (1996b). Mutational analysis of the human nucleotide excision repair gene ERCC1. Nucleic Acids Res. *24*, 3370–3380.

Simon, G.R., Sharma, S., Cantor, A., Smith, P., and Bepler, G. (2005). ERCC1 expression is a predictor of survival in resected patients with non-small cell lung cancer. Chest *127*, 978–983.

Singh, S., Folkers, G.E., Bonvin, A.M., Boelens, R., Wechselberger, R., Niztayev, A., and Kaptein, R. (2002). Solution structure and DNA-binding properties of the C-terminal domain of UvrC from *E. coli*. EMBO J. *21*, 6257–6266.

Tian, M., Shinkura, R., Shinkura, N., and Alt, F.W. (2004). Growth retardation, early death, and DNA repair defects in mice deficient for the nucleotide excision repair enzyme XPF. Mol. Cell. Biol. *24*, 1200–1205.

Volker, M., Mone, M.J., Karmakar, P., van Hoffen, A., Schul, W., Vermeulen, W., Hoeijmakers, J.H., van Driel, R., van Zeeland, A.A., and Mullenders, L.H. (2001). Sequential assembly of the nucleotide excision repair factors in vivo. Mol. Cell *8*, 213–224.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. J. Mol. Graph. *8*, 52–56.

Weeda, G., Donker, I., de Wit, J., Morreau, H., Janssens, R., Vissers, C.J., Nigg, A., van Steeg, H., Bootsma, D., and Hoeijmakers, J.H. (1997). Disruption of mouse ERCC1 results in a novel repair syndrome with growth failure, nuclear abnormalities and senescence. Curr. Biol. *7*, 427–439.

Zhu, X.D., Niedernhofer, L., Kuster, B., Mann, M., Hoeijmakers, J.H., and de Lange, T. (2003). ERCC1/XPF removes the 3′ overhang from uncapped telomeres and represses formation of telomeric DNA-containing double minute chromosomes. Mol. Cell *12*, 1489–1498.

Accession Numbers

Coordinates have been deposited in the RCSB Protein Data Bank with accession code 1Z00, and the chemical shifts have been deposited in BMRB with accession number 6551.

Published online 24 August 2007 *Nucleic Acids Research, 2007, Vol. 35, No. 17* ***5789–5798***
doi:10.1093/nar/gkm503

Analysis of the XPA and ssDNA-binding surfaces on the central domain of human ERCC1 reveals evidence for subfunctionalization

Konstantinos Tripsianes, Gert E. Folkers, Chao Zheng, Devashish Das, Jeffrey S. Grinstead, Robert Kaptein and Rolf Boelens*

Department of NMR spectroscopy, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

Received May 11, 2007; Revised June 6, 2007; Accepted June 10, 2007

ABSTRACT

Human ERCC1/XPF is a structure-specific endonuclease involved in multiple DNA repair pathways. We present the solution structure of the non-catalytic ERCC1 central domain. Although this domain shows structural homology with the catalytically active XPF nuclease domain, functional investigation reveals a completely distinct function for the ERCC1 central domain by performing interactions with both XPA and single-stranded DNA. These interactions are non-competitive and can occur simultaneously through distinct interaction surfaces. Interestingly, the XPA binding by ERCC1 and the catalytic function of XPF are dependent on a structurally homologous region of the two proteins. Although these regions are strictly conserved in each protein family, amino acid composition and surface characteristics are distinct. We discuss the possibility that after XPF gene duplication, the redundant ERCC1 central domain acquired novel functions, thereby increasing the fidelity of eukaryotic DNA repair.

INTRODUCTION

During DNA replication, recombination and repair, double-stranded DNA inevitably forms three- or four-way junctions, bubbles, flaps or broken ends with single-stranded extensions. These irregular structures must be processed correctly in order to successfully complete DNA metabolism and thereby maintain genome integrity. This task is accomplished by structure-specific endonucleases specialized in pruning downstream of branch, flap or bubble structures by incision at junctions between double- and single-stranded DNA (1). Inactivation or malfunctioning of these enzymes causes genetic defects or cancer, underlying their importance in genome stability.

A remarkable class of structure-specific endonucleases in humans is XPF. The protein family is characterized by the presence of the ERCC4 domain and consists of seven members (XPF, MUS81, ERCC1, EME1, EME2, FANCM and FAAP24) (2). Only XPF and MUS81 have nuclease activity, which is mediated by the conserved core nuclease motif ($ERKX_3D$) (3–5). Their catalytic function depends on heterodimer formation with the non-catalytic family members. XPF forms an obligate complex with ERCC1 and functions primarily in nucleotide excision repair (NER), a versatile pathway able to detect and remove a variety of DNA lesions induced by UV light and environmental carcinogens. The ERCC1/XPF heterodimer has additional roles in DNA interstrand cross-link (ICL) repair (6) and telomere maintenance (7). The symptoms of the first patient with inherited ERCC1 deficiency (8) and of a patient with a novel XPF mutation (9) are distinct from the classical NER phenotype, and underscore the pleiotropic function of ERCC1/XPF.

In contrast to eukaryotes, archaea have a single homolog of the XPF endonuclease that forms homodimers. The archaeal XPF minimally consists of the catalytic nuclease domain followed by a DNA-binding domain containing two consecutive helix-hairpin-helix motifs (HhH_2 domain). Dimerization occurs between both the nuclease and HhH_2 domains of each subunit (Figure 1A) (3). The strong preference of the archaeal endonucleases for 3′ flap DNA substrates is explained by the function of their individual domains. Structural data suggest a model for DNA binding where one HhH_2 domain of the dimer binds to an upstream and the other to a downstream DNA duplex, sharply bending the DNA substrate and thereby allowing one active site of the dimeric nuclease domain to cleave the 3′ protruding single strand (10). Combination of mutations in nuclease and HhH_2 domains that do not disrupt the dimeric interfaces but impair the individual functions (DNA cleavage or DNA binding), provide evidence for this model (11).

*To whom correspondence should be addressed. Tel: +31 30 2534035; Fax: +31 30 2537623; Email: r.boelens@chem.uu.nl

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

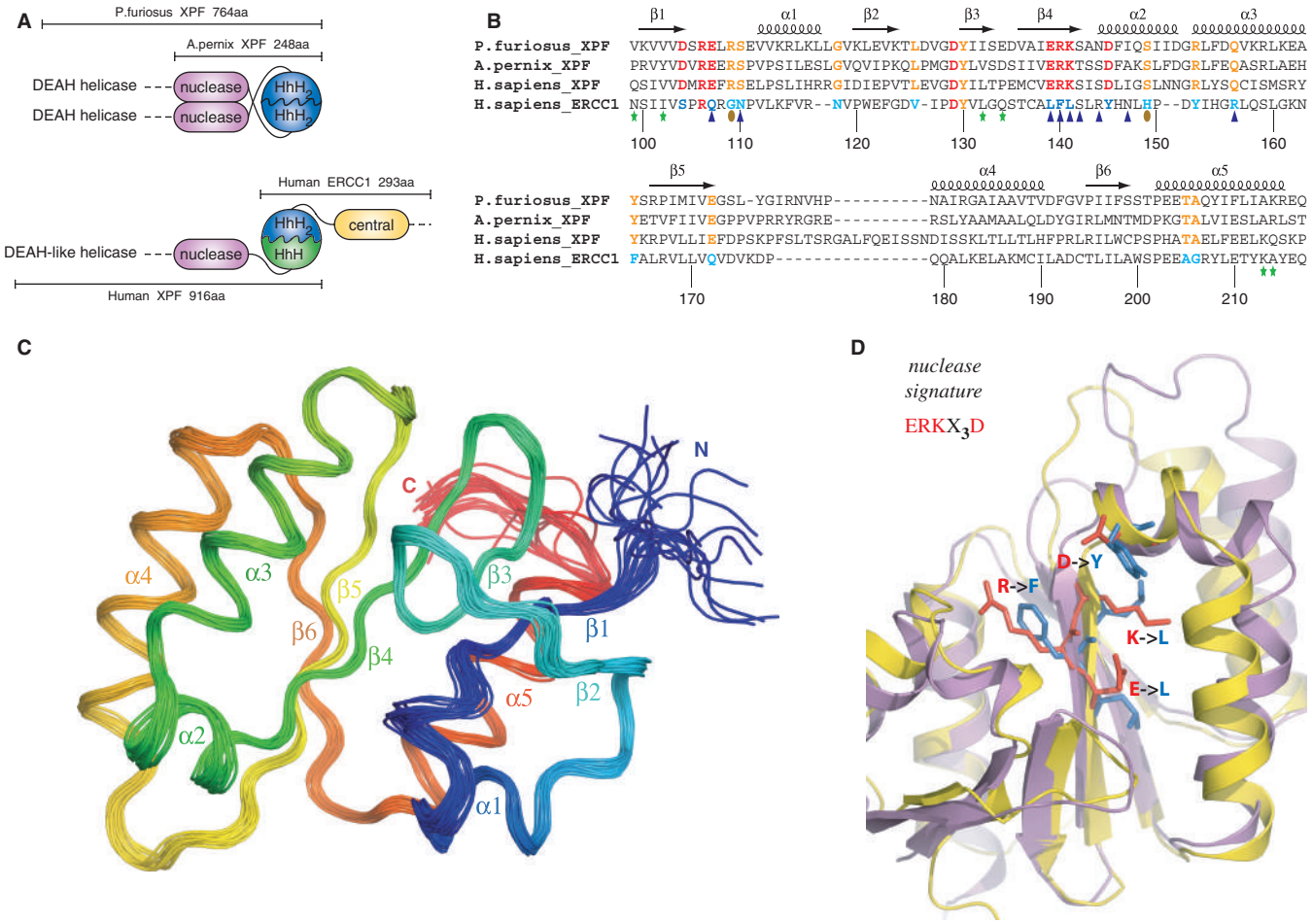


Figure 1. (**A**) Domain organization of the archaeal XPF homodimeric members and the human ERCC1/XPF heterodimer. (**B**) Structure-based sequence alignment of the nuclease XPF domains from archaea to human and the corresponding central domain of human ERCC1. Secondary structure elements of the prototype XPF nuclease fold are indicated at the top of the sequences. Catalytic residues in the XPF nucleases are colored red and their corresponding substitutions in ERCC1 blue. Other invariant residues in XPF domains are depicted in orange, while the ERCC1 equivalents are depicted in cyan. Residues of cERCC1 perturbed largely upon XPA titration are indicated by blue triangles and those appear only in the final complex by brown ellipses. Green asterisks indicate cERCC1 residues perturbed by DNA titration. cERCC1 sequence is numbered at the bottom. (**C**) Ensemble of the final 20 structural conformers of cERCC1 as determined by solution NMR. Secondary structure elements and N- and C-termini are labeled. (**D**) Superposition of the crystal XPF nuclease structure (2bgw) from *A. pernix* (purple) and the solution NMR structure (2jpd) of human cERCC1 (yellow). Emphasis is given to the nuclease signature and the corresponding substitutions in cERCC1.

The DNA substrate specificity is different for the human ERCC1/XPF heterodimer. The human enzyme is known to specifically incise hairpin, bubble or splayed-arm DNA substrates (12,13), which consist of only one duplex (upstream). Previous data indicated the role of the C-terminal ERCC1 'canonical' HhH_2 domain in engaging the upstream DNA duplex of a hairpin substrate (14). This function requires heterodimerization with the C-terminal HhH domain of the human XPF partner to form a complex analogous to the archaeal HhH_2 dimer interface (12,14). Previous structural data (12) and the data presented here illustrate the structural similarity between the human ERCC1 central domain and the archaeal XPF nuclease domain. Unlike the archaeal nuclease dimer interface, there is no evidence that the corresponding domains of human ERCC1 and XPF interact with each other in solution (Figure 1A) (12).

The catalytic function of the ERCC1/XPF endonuclease is crucial for NER. NER operates through a 'cut and patch' mechanism by excising and removing a short stretch of DNA containing the lesion, and subsequently restoring the genetic information by repair synthesis using the undamaged strand as the template. The incision step involves the sequential and coordinated assembly of the DNA damage sensor XPC-HR23B, the transcription/repair factor TFIIH, the architectural protein XPA, the ubiquitous ssDNA-binding protein RPA, and the two structure-specific endonucleases ERCC1/XPF and XPG, responsible for the incisions 5′ and 3′ to the damaged site, respectively (15). A main role in the progress of the reaction is attributed to XPA, which in complex with RPA probes the DNA helix conformations (16) and participates in multiple interactions with the other NER factors (15). XPA is required for the recruitment of ERCC1/XPF in the

NER pre-incision complex (17,18). Consistent with this regulatory function, *in vitro* studies with recombinant proteins and *in vivo* studies using a yeast two-hybrid system have demonstrated interactions between XPA and ERCC1 (19).

Here, we report the solution structure of the ERCC1 central domain (cERCC1) and investigate its interactions with XPA and DNA at the molecular level. Using biochemical studies and NMR titration experiments, we have identified two distinct interaction surfaces of cERCC1 that mediate XPA and DNA binding. Interestingly, the two interactions can take place simultaneously to yield a cERCC1/DNA/XPA ternary complex, which in turn explains the important role of ERCC1 in targeting its catalytic XPF partner to the NER pre-incision complex.

MATERIALS AND METHODS

Cloning, protein expression and protein purification

ERCC1 central domain (cERCC1, residues 96 to 219) was PCR-amplified from a vector containing the full-length ERCC1 gene and subcloned into a pET28b (Novagen) expression vector between BamHI and XhoI sites. Downstream of the XhoI site a linker together with a his-tag has been engineered. XPA constructs were derived from a full-length XPA vector purchased from RZPD (clone ID IRAUp969B1273D6) and subcloned into a pLICHIS vector using the enzyme free cloning (EFC) strategy (20). The same procedure was followed for the GST fusion proteins, either XPA or cERCC1 (pLICHISGST vector). As a result all the EFC constructs contain an N-terminal his-tag, while the cERCC1 construct bears a C-terminal his-tag. The nucleotide sequences of the cloned DNAs were confirmed by sequencing. The cERCC1-HIS, HIS-XPA, HIS-GST-XPA and HIS-GST-cERCC1 proteins were expressed in BL21(DE3) Rosetta cells (Novagen) and were subject to two-step purification as has been described (21). When appropriate, HIS-GST-cERCC1 was cleaved with thrombin in the presence of 1 mM Ca^{2+}, 50 mM Tris (pH 8.0) and 100 mM NaCl in order to obtain the untagged protein. Finally, the protein samples were exchanged either to 50 mM sodium phosphate (pH 5.5), 100 mM NaCl for the cERCC1 structure determination, or to 50 mM Tris (pH 8.0), 100 mM NaCl for the binding studies.

NMR spectroscopy

Multidimensional NMR experiments were carried out at 290 K on Bruker AVANCE 600 and 900 NMR spectrometers equipped with TXI triple-resonance probes in 50 mM sodium phosphate (pH 5.5), 100 mM NaCl by using the cERCC1-HIS protein. Spectra were processed using the NMRPipe software package (22) and analyzed with Sparky (23). The ^{1}H, ^{15}N and ^{13}C resonance assignments are 97% complete and were made using standard triple resonance techniques, 3D NOESY-(^{15}N, ^{1}H)-HSQC spectra, and 3D NOESY-(^{13}C, ^{1}H)-HSQC spectra (both with a mixing time of 60 ms) (24). The chemical shifts have been deposited at the BMRB with accession number 15240.

Structure calculations

Automatic NOE assignment and structure calculations were performed using the program CYANA version 2.0 (25). Hydrogen bond restraints were defined when they were consistent with the secondary chemical shift data and expected NOE contacts and only for the helical parts of the protein. The set of 2669 NOE restraints determined by CYANA, together with restraints for 31 hydrogen bonds and ϕ and ψ torsion angle restraints (174) derived from TALOS (26) were used in a water refinement run according to the standard RECOORD protocol (27) utilizing CNS (28). Molecular images were generated with PyMol (29). Coordinates have been deposited at the PDB with accession code 2jpd and the structural restraints at the BMRB with entry number 15240.

EMSAs

All the DNA substrates were 5′ ^{32}P-labeled with T4 polynucleotide kinase and purified on a polyacrylamide gel under native conditions. The nucleotide sequences for ss20, b10 and ds30 have been described before (30). Per reaction, 100 fmol of substrate was incubated with the appropriate amount of ERCC1 protein in binding buffer (50 mM Tris, pH 8.0, 100 mM NaCl, 10% glycerol and 1 mg/ml BSA). After incubation for 30 min at 4°C, samples were loaded onto 7.5% native polyacrylamide gels containing 0.5× TBE and run at 4°C. Alternatively Electrophoretic mobility shift assay (EMSA) reactions were loaded on a 3.5% agarose gel containing 0.5% TBE, run at 4 or 20°C yielding essentially identical results. Gels were visualized and quantified using a phosphor imager (BioRad) as described before (30). For depletion experiments, 20 µl of MagneHis beads (Promega) or GST agarose beads (Sigma), were extensively washed with binding buffer, and after buffer removal the EMSA reaction was added to the beads. After 15 min incubation with regular mixing, the beads were removed from the EMSA reaction using a magnet or centrifugation and the reaction mixture was loaded on acrylamide gel or, when appropriate, the substrate was added to the depleted reaction mixture.

GST pull-down assay

GST pull-down assay was performed and quantified as described before (31) using the indicated cERCC1 concentrations and 3–5 µg of GST proteins in 150 µl of 50 mM Tris, 100 mM NaCl, 10% glycerol, 1 mM DTT, 0.2 mM PMSF and 20 mg/ml of BSA (pH 8.0). The semi-quantitative experiments were performed in 50, 150, 500, 1500 and 5000 µl buffer with a constant amount of cERCC1 and GST-XPA.

NMR titrations

cERCC1 titrations with XPA and DNA were performed on a Bruker AVANCE 700 NMR spectrometer and were monitored with 2D ^{1}H-^{15}N HSQC experiments. In all

Table 1. Structural statistics of the structure ensemble of human cERCC1 (residues 96–219)

Rmsd (Å) with respect to mean[a]	
Backbone/Heavy atoms	$0.40 \pm 0.04/0.89 \pm 0.08$
Number of experimental restraints	
Intraresidue NOEs	632
Sequential NOEs ($\|i-j\| = 1$)	678
Medium range NOEs ($1<\|i-j\|<4$)	479
Long-range NOEs ($\|i-j\|>4$)	880
Total NOEs	2669
Dihedral angle restrains	174
Hydrogen bonds	31
Restraint violations	
NOE distances with violations >0.3 Å	0.00 ± 0.00
Dihedrals with violations >3°	0.00 ± 0.00
Rmsd for experimental restraints	
All distance restraints (2669) (Å)	0.013 ± 0.005
Torsion angles (174) (°)	0.579 ± 0.077
CNS energies after water refinement	
E_{vdw} (kcal/mol)	-614 ± 17
E_{elec} (kcal/mol)	-4914 ± 61
Rmsd from idealized covalent geometry	
Bonds (Å)	0.01 ± 0.00
Angles (°)	1.15 ± 0.03
Impropers (°)	1.34 ± 0.06
Ramachandran analysis	
Residues in the favored regions (%)	90.68 ± 2.04
Residues in additional allowed regions (%)	8.16 ± 1.71
Residues in generously allowed regions (%)	0.81 ± 0.65
Residues in disallowed regions (%)	0.36 ± 0.68

[a]Residues 101–215 of cERCC1.

cases, the concentration of cERCC1 was 0.2 mM in 50 mM Tris (pH 8.0), 100 mM NaCl. XPA and DNA were dissolved in the same buffer and therefore salt and pH did not vary throughout the experiments. DNA binding was feasible only with the cERCC1 construct lacking the C-terminal his-tag, while XPA binding was identical, regardless of the his-tag presence. Normalized chemical shift changes were calculated by using the equation: $\delta = ([\delta_{HN}]^2 + [\delta_N/6]^2)^{0.5}$.

RESULTS AND DISCUSSION

Structure analysis

We have determined the solution structure of the ERCC1 central domain (cERCC1) to elucidate the molecular details of its proposed functions. We have collected a large number of distance and dihedral angle restraints that yielded an ensemble of 20 conformers with very good convergence (Figure 1C). The quality of the structure can be judged by the summary of the structural and restraint statistics given in Table 1. All the secondary structure elements are well defined including the short loops that join them. cERCC1 comprises a compact architecture and folds as a six-stranded β-sheet flanked by five α-helices on both sides. As described before (12), this fold is reminiscent of the type II restriction endonucleases, to which the catalytic domain of XPF also belongs.

Overall, the solution NMR structure of cERCC1 is in very good agreement with the crystal structure of the same domain (2a1i) (rmsd 1.1 Å for 108 C_α atoms). The only noticeable difference relates to the last few residues of helix α3 and the loop connecting this structure element with β5. The presence of a mercury atom (linked to C137) in the crystal structure may account for the substantial side-chain rearrangements within this loop. When compared to the crystal structures of the archaeal nucleases, helices α1, α2, and strands β1, β2 appear to be shorter in the solution structure of cERCC1. Solution cERCC1 and the crystal structures of archaeal XPF nucleases (2bgw and 1j23) contain the same number of secondary structure elements arranged in a very similar manner (rmsd 1.8 and 2.0 Å for *Aeropyrum pernix* and *Pyrococcus furiosus* nucleases, respectively, for 108 C_α atoms), although in both cases the primary sequence homology is very low (Figure 1B and D).

Remarkably, conserved residues of XPF nuclease scattered in the primary sequence superimpose structurally with residues conserved in the ERCC1 sequence family (Figures 1 and 3). Whereas the conservation in XPF is directly related to the catalytic function, ERCC1 has preserved the same fold but lacks the essential residues for catalysis. The fold similarities, coupled with the obligate nature of the heterodimerization (32), are in full agreement with the common origin of the two proteins (33).

Minimal domains for ERCC1–XPA interaction

Truncation studies of ERCC1 have mapped the XPA interaction site to a region between residues 91 and 118 (34). From the cERCC1 structure and its compact fold, we predict that such truncations will have a devastating effect on the structural integrity of the central domain (96–219). For XPA, the ERCC1 interaction region seems to be located in a small stretch containing two highly conserved motifs rich in glycine and glutamic acid residues (72-GGGFILEEEEEEE-84, conserved residues are underlined) (19). We have prepared a HIS-GST-XPA construct (59–99) containing the conserved stretch, and examined its ability to interact with cERCC1 in a GST pull-down assay. Indeed, the GST XPA fusion protein was able to specifically bind cERCC1 (Figure 2A), confirming that this domain of ERCC1 is sufficient for interactions with XPA. No binding of cERCC1 to GST bound agarose beads or uncharged agarose beads was observed (Figure 2A, data not shown). This XPA peptide is unstructured (35) and appeared very sensitive to proteolysis both during overexpression in *Escherichia coli* and as pure protein after extensive purification (Figure 2A and B). We next performed a semi-quantitative GST pull-down assay (31) and estimated an apparent K_d of 1 μM. Since we reached only 20% binding saturation at the two highest protein concentrations (3 and 10 μM of cERCC1) under these experimental conditions, we could not determine the binding constant more accurately (Figure 2B).

Functional disparity of the XPF nuclease and ERCC1 central domains

The cERCC1–XPA interaction specificity has been explored using NMR titration experiments. Titration with the HIS-XPA peptide causes extensive specific

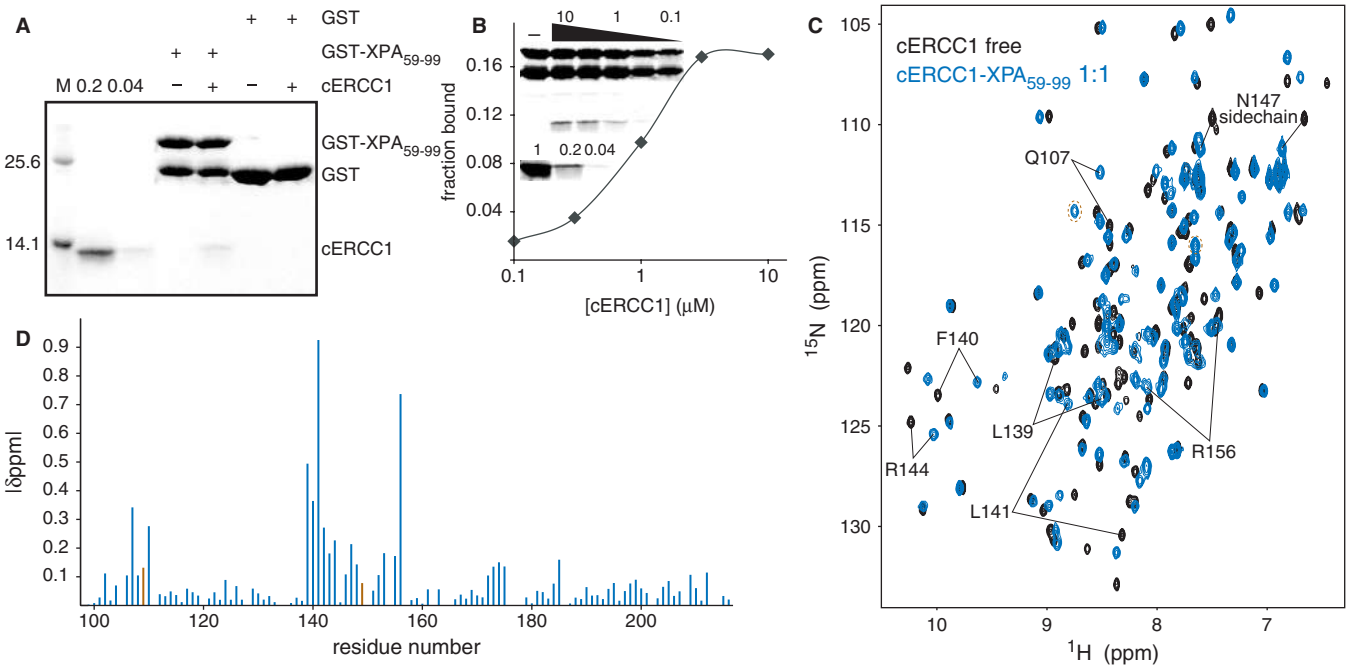


Figure 2. ERCC1–XPA interactions. (**A**) GST pull-down assay with 3 μg of GST or GST-XPA fusion proteins in the presence or absence of 2 μM cERCC1. Here, 0.2 and 0.04 refer to respectively 20 and 4% of the input cERCC1 protein present in the GST pull-down assay. (**B**) Semi quantitative GST pull-down assay showing the fraction of cERCC1 bound to 3 μg of GST-XPA at the indicated [cERCC1] (μM). The upper part of inset shows a representative GST pull-down assay, where the various cERCC1 concentrations used are depicted above (10, 3, 1, 0.3 and 0.1 μM). The lower panel shows respectively 100, 20 and 4% of cERCC1 added to the assay. (**C**) Chemical shift perturbation of the cERCC1 ^{1}H-^{15}N HSQC upon complex formation with XPA. Free cERCC1 spectrum is in black and XPA-bound spectrum in blue, while brown circles in the bound spectrum indicate G109 and H149 resonances. (**D**) Normalized chemical shift changes between free and XPA-bound forms versus the cERCC1 sequence. The p.p.m. difference for G109 and H149 was calculated by their resonances in the free cERCC1 spectrum at pH 5.5, and are colored brown.

changes in the cERCC1 ^{1}H-^{15}N HSQC NMR spectra. The titrations show slow exchange regime with respect to the NMR chemical shift timescale, in agreement with the determined affinity (Figure 2C). We have been able to assign all the amide resonances of cERCC1 in the bound form (1:1 complex stoichiometry) by using 3D NOESY-(^{15}N, ^{1}H)-HSQC spectra. The pattern of the backbone amide NOEs does not change substantially compared to the free cERCC1 protein. This indicates that the overall structure is maintained, and led us to conclude that the observed chemical shift perturbations arise from specific contacts with the XPA peptide.

The most significant perturbations map to the surface of the V-shaped groove of the cERCC1 structure and involve many residues that compose the corresponding DNA cleavage site of the archaeal XPF endonucleases (Figures 2D and 3). Interestingly, among the largest chemical shifts of cERCC1 upon XPA binding are those of L139, F140 and L141. These ERCC1 residues correspond to the characteristic strongly conserved ERK catalytic triad of the XPF nuclease domain (Figure 1). The rest of the significantly perturbed residues are polar or positively charged (Q107, N110, S142, R144, N147 and R156), are competent for hydrogen bonding with the XPA peptide and are more conserved than the surrounding regions of the cERCC1 V-shaped groove (Figure 3). Additionally, of special interest are the amide signals of G109 and H149, which are not observed in the free protein spectrum due to fast exchange with the solvent at pH 8 (these amide resonances are observable at pH 5.5). However, both give sharp signals at the end of the titration, an indication that the solvent exchange in the free protein is quenched in the complex. These residues reside at the rim of the V-shaped binding site and may be involved in hydrogen bonding with the XPA peptide (Figure 3).

Although ERCC1 residues forming the XPA-binding site are different from the acidic and basic residues which form the active site in XPF (5), there are no significant changes in the backbone fold and the spatial side-chain orientations in the structure remain unaffected. However, their different chemical properties result in a distinct charge and hydrophobicity distribution on the structure surface (Figure 3). While XPF displays negative charge at the DNA cleavage site (important for divalent cation coordination), the equivalent XPA-binding site in ERCC1 is neutral or slightly positively charged. Therefore, the ERCC1–XPA interaction is a combination of hydrophobic and electrostatic interactions possibly involving the glutamic acid stretch of XPA (19) and two conserved positively charged residues (R106 and R156) of cERCC1.

DNA binding by the ERCC1 central domain

Given the proposed common origin of ERCC1 and XPF (33), it can be expected that the degenerated catalytic site of ERCC1 could still bind DNA in a similar fashion as the archaeal XPF active site (12). Because our ERCC1–XPA

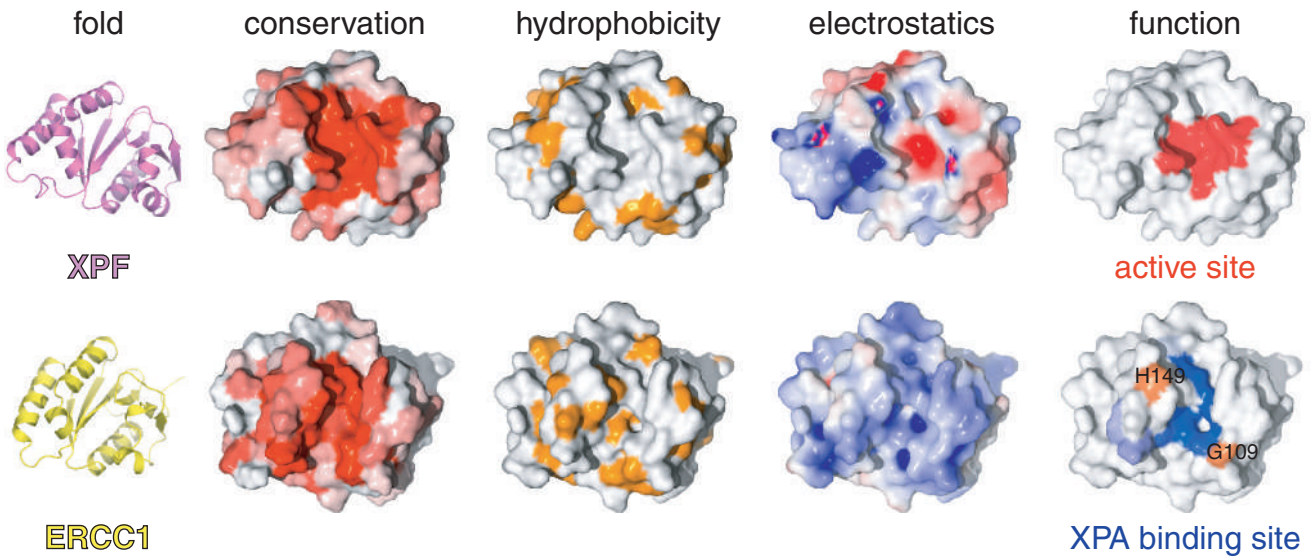


Figure 3. Common fold, different properties and distinct functions for XPF and ERCC1. Sequence conservation either for XPF or ERCC1 proteins is colored from white (non-conserved) to red (highly conserved). The opposite face has no significant conservation for either protein. Hydrophobic side chains are colored orange. Electrostatic surface potentials were calculated using APBS (39) and colored blue for positive or red for negative charge potential. Active site residues of XPF were chosen based on previous mutational studies (5) and the XPA-binding site of ERCC1 was identified by our NMR titrations.

interaction data render this unlikely, we explored whether cERCC1 is able to bind DNA as has been suggested previously (12).

We performed an electrophoretic mobility shift assay (EMSA) where increasing amounts of cERCC1-HIS protein were added to ssDNA. While we saw a loss of free probe, we failed to see any complex formation under these conditions. If the same protein–DNA complex was loaded on a 3.5% agarose gel, we detected a weak smeary complex at the highest protein concentration (Figure 4A). However, by using a HIS-GST-cERCC1 protein, we observed formation of the complex with ssDNA (Figure 4B). This suggests that the C-terminal his-tag interferes with DNA binding. Therefore, we used the HIS-GST-cERCC1 protein, which we cleaved with thrombin to obtain the untagged protein. DNA binding with the untagged cERCC1 demonstrated a faster migrating complex with various ssDNA substrates (data not shown). HIS-GST-cERCC1/DNA complex formation is prevented by incubation of the reaction mixture with MagneHIS magnetic beads or glutathione agarose beads, which depletes HIS-GST-cERCC1 from the reaction mixture either prior to or after the addition of DNA. If beads were added after complex formation, a small but significant decrease in the amount of free probe was observed. Depletion of the HIS-GST tagged cERCC1 protein from the binding reaction confirms that the protein–DNA complex is indeed formed by the HIS-GST-cERCC1 protein (Figure 4C).

In agreement with earlier studies, under these conditions no binding was observed for dsDNA (12), while fairly comparable binding affinities were found for both ssDNA and bubble substrates with 10 or 20 unpaired bases. We calculated an apparent K_d of $2.5 \pm 0.7\,\mu M$ for the HIS-GST-cERCC1 protein bound to a bubble10 (b10) substrate in the presence of 100 mM NaCl (Figure 4B). Corroborating our results, an equilibrium binding titration experiment using fluorescence anisotropy has demonstrated ssDNA binding for the same domain with an apparent K_d of 10 μM (12). The relatively small difference in affinity can be well explained by different salt conditions or substrate used.

For the NMR titrations we performed a thrombin cleavage on HIS-GST-cERCC1, to remove the his-tag. The untagged cERCC1 ^{1}H-^{15}N HSQC spectrum is identical to that of cERCC1-HIS, except for the two amides preceding the artificial his-tag (Supplementary Figure 1). Addition of the b10 substrate to this cERCC1 protein caused a limited number of specific perturbations that were unambiguously assigned. Again, the perturbed resonances exhibit slow exchange behavior in the NMR titrations and indicate contacts with the DNA. The affected resonances include the backbone amides of N99, I102, L132, K213, A214 and the side-chain amide of Q134 (Figure 4D and E). Compared to the free-protein spectrum we only miss T211, which cannot be identified. The perturbation induced by the b10 DNA was complete at equimolar concentration with cERCC1. Chemical shift mapping reveals that all perturbations are in the vicinity of the C-terminus of the cERCC1 construct, consistent with the presence of a flexible tag interfering with DNA binding. The DNA-binding region resides at the surface of the structure where the N- and C-termini meet (Figure 4F). The perturbed residues belong to well-defined structure elements. The only positively charged residue we identified (K213) is fully conserved in the ERCC1 proteins and may explain the dependence of DNA binding on salt concentration. All the other affected residues are less conserved than the residues involved in XPA binding. The composition of the affected residues together with the

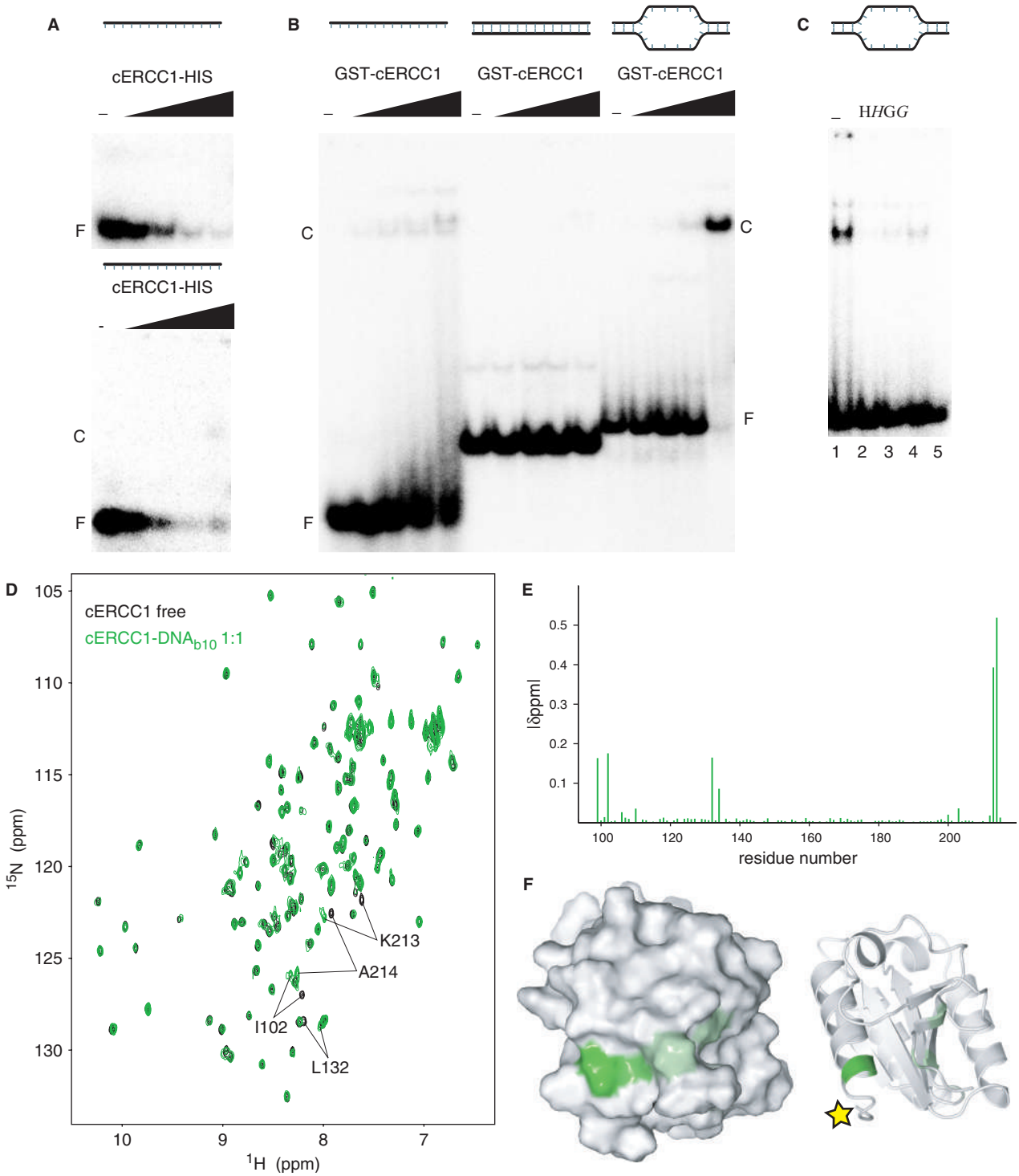


Figure 4. DNA binding by cERCC1. (**A**) EMSA with a 20-mer ssDNA substrate using increasing concentrations of cERCC1-HIS protein (0, 0.04, 0.2, 1, and 5 μM), loaded on a 7.5% acrylamide gel (upper panel) or a 3.5% agarose gel (lower panel). (**B**) EMSA using ssDNA (20-mer), dsDNA (30 bp) and bubble10, a dsDNA with 10 unpaired bases as substrate, in the presence of 0, 0.125, 0.25, 0.5 and 1 μM HIS-GST-cERCC1. (**C**) Inhibition of cERCC1–b10 DNA complex formation ([cERCC1] = 0.5 μM) by depletion of HIS-GST-cERCC1 from the EMSA reaction by the addition of MagneHis beads (H) or GST agarose beads (G) prior to or after addition of DNA (italic). F: free DNA, C: cERCC1–DNA complex. (**D**) Chemical shift perturbation after addition of b10 DNA. The spectrum of the free protein is shown in black and the spectrum after addition of equimolar amount of b10 in green. (**E**) Normalized chemical shift changes between free and b10-bound forms versus the cERCC1 sequence. (**F**) Surface and cartoon representations of cERCC1 colored according to normalized chemical shifts. Yellow star indicates the position of the C-terminal his-tag that was interfering with DNA binding.

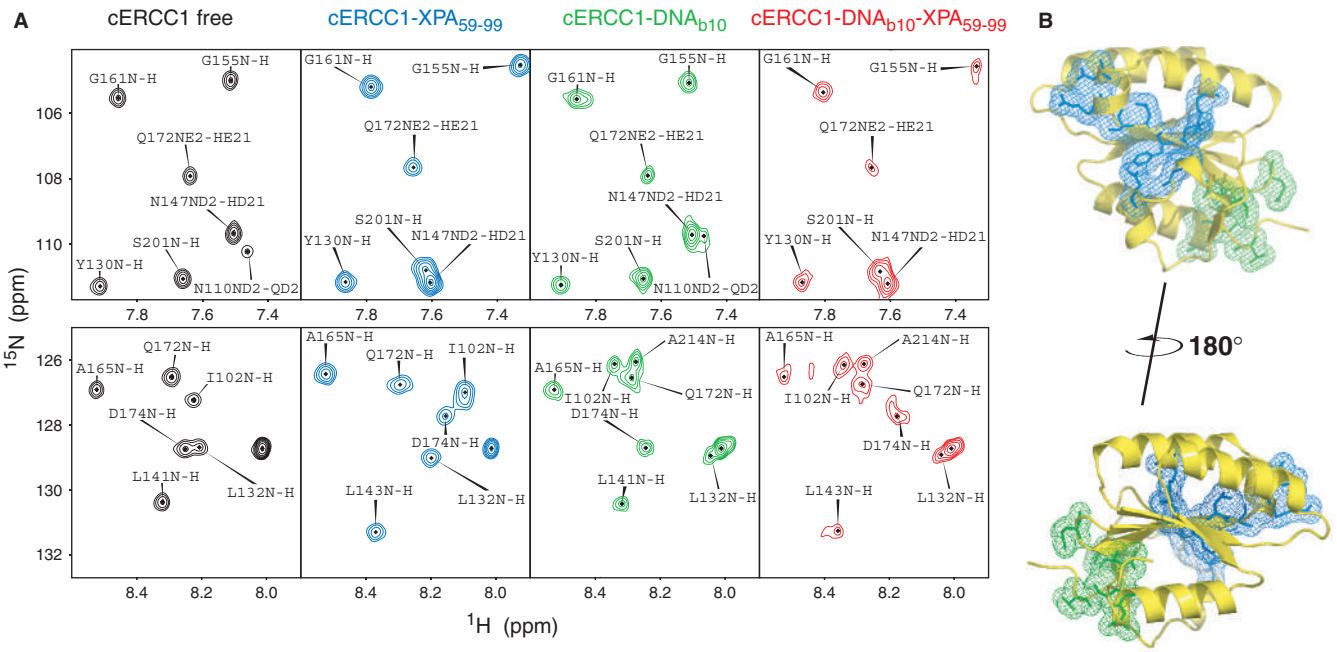


Figure 5. Dual function of cERCC1. (**A**) Indicative portions of the ^{1}H-^{15}N HSQC spectrum for the free cERCC1, bound to either XPA (1:1) or DNA (1:1) separately, and the ternary complex (1:1:1). (**B**) Views of the two binding sites on cERCC1 structure.

slow exchange behavior suggests a large contribution to the binding by hydrophobic interactions with the DNA bases.

Summarizing, both the biochemical and the NMR experiments show that the ERCC1 central domain binds to DNA. The DNA-binding site we identified on cERCC1 by the NMR titrations suggests that cERCC1 contacts only a small part of the ssDNA, probably three to four bases. The situation is different in the full-length heterodimer, where other DNA-binding surfaces, such as previously established for the C-terminal ERCC1/XPF domains (12,14), assist in DNA recognition.

cERCC1 can bind simultaneously to XPA and ssDNA

Chemical shift perturbation experiments revealed distinct binding surfaces for ssDNA and XPA, and therefore, it is probable that both can bind simultaneously to cERCC1. We performed EMSA experiments in the presence of an equimolar or 5-fold excess of the HIS-XPA peptide. Addition of XPA did not influence binding of cERCC1 to ssDNA, while the binding affinity of XPA for the cERCC1 domain under these conditions would suggest that the vast majority of cERCC1 is in complex with XPA. We did not observe formation of a slower migrating HIS-GST-cERCC1/ssDNA/XPA super complex, possibly because the increase in mass by the addition of the HIS-XPA peptide was too small to discern (data not shown).

To independently confirm the formation of a ternary cERCC1/ssDNA/XPA complex, we added the HIS-XPA peptide to the cERCC1-bubble10 protein–DNA complex and followed the chemical shift perturbations in the ^{1}H-^{15}N HSQC spectra. As shown in Figure 5, addition of an equimolar amount of the XPA peptide causes the same amide displacements as in the titration with the free protein. On the contrary, most amides influenced by the DNA binding remain unaffected by the presence of XPA (Figure 5A). I102 and N110 were perturbed by both XPA and b10 DNA when the titrations were performed independently. In the ternary complex, however, I102 remains close to the position as in the DNA bound form. Conversely, the side-chain amide of N110, although affected by the DNA, adopts the distinct XPA bound position upon XPA addition. These data strongly indicate the formation of the ternary ERCC1/DNA/XPA complex. DNA and XPA-binding sites are distinct, and the two interactions can happen simultaneously (Figure 5B).

Functional role of the ERCC1 central domain in the heterodimer

Since the catalytic activity has been preserved in the human XPF protein through strong conservation of the nuclease signature (5), the human nuclease fold should be identical to that of the archaeal counterparts. In that sense, human XPF nuclease and human ERCC1 central domains are expected to exhibit the same fold. Moreover, the ERCC1 and XPF HhH_2 domains feature a common architecture and come into tight association exactly in the same way that the HhH_2 domains of both archaeal species do (10,11,14). Therefore, in structural terms the human ERCC1 and XPF proteins share the essential architectural subunits observed in the short XPF homodimer of *A. pernix*. The structural similarities within the XPF family provide additional evidence for the proposed common origin of the human ERCC1 and XPF proteins (33). Because ERCC1 is absent in archaea, this gene is thought to have been acquired from an ancient XPF gene

duplication in the eukaryal lineage. The ERCC1 and XPF genes have subsequently evolved by the process called subfunctionalization (36,37). This model suggests that after gene duplication, both copies may be reciprocally preserved through the fixation of complementary loss-of-subfunction mutations, which results in a partitioning of the tasks of the ancestral gene. From the functions present in the ancestral XPF protein (archaea), human ERCC1 retained the canonical HhH_2 domain that acts as a DNA-binding domain (14), while XPF retained the catalytic activity (5). Once the separation of the ancestral subfunctions occurred in ERCC1 and XPF genes, only the heterodimeric protein complex could restore the original function. Additionally, due to genetic alterations, the second helix-hairpin-helix motif of human XPF degenerated, yet the fold remained crucial for stabilizing the corresponding intact domain of human ERCC1 (14). Similarly, the central domain of ERCC1 lost the catalytic activity by sequence drift, but despite adoption of a novel dual function (XPA interaction and DNA binding), the 3D fold was preserved.

We suggest a mechanistic model for the heterodimeric function based on the model for the homodimeric archaeal homolog (Figure 6) (10,38). Human ERCC1 corresponds to the archaeal protomer that binds with the HhH_2 domain to the upstream duplex. Human XPF corresponds to the archaeal protomer that recognizes the downstream duplex. Since it cannot make contacts with the minor groove (14), the specificity of the human heterodimer has shifted to splayed-arms substrates consisting of only one duplex. This does not exclude contribution of the XPF HhH domain to ssDNA interactions, as reported before (12). We show here that the central domain of ERCC1 is also involved in ssDNA binding. Additional DNA interactions are likely for the nuclease of XPF, analogous to the archaeal case (38). Furthermore, ssDNA binding by cERCC1 will be stimulated by specific interactions with XPA, already present in the NER pre-incision complex and possibly bound to the DNA via its own DNA-binding site. Most importantly, through these coordinated interactions of the ERCC1 domains the XPF protein will be positioned to cleave the 3′ protruding strand (limon), thereby retaining the polarity present in the archaeal homodimer.

Our model underlines the significant role of ERCC1 in the context of the full-length heterodimer. XPF is the catalytic module but the ERCC1 domains guarantee that the enzymatic activity is targeted properly. The presence of multiple distinct DNA-binding surfaces within the ERCC1/XPF/XPA/RPA repair protein intermediate coordinates cleavage to occur only when the DNA damage is recognized correctly by the NER machinery. The duplication of the ancestral XPF gene within the eukaryal kingdom resulted in an obligate heterodimer through loss of function (with an altered HhH_2 domain of XPF and a degenerated catalytic domain of ERCC1) and adoption of novel functions (ssDNA and XPA binding of cERCC1). This permitted additional quality control mechanisms through a more complicated molecular interaction network, mediated by the novel functional domains, and thereby improving the fidelity of DNA damage repair.

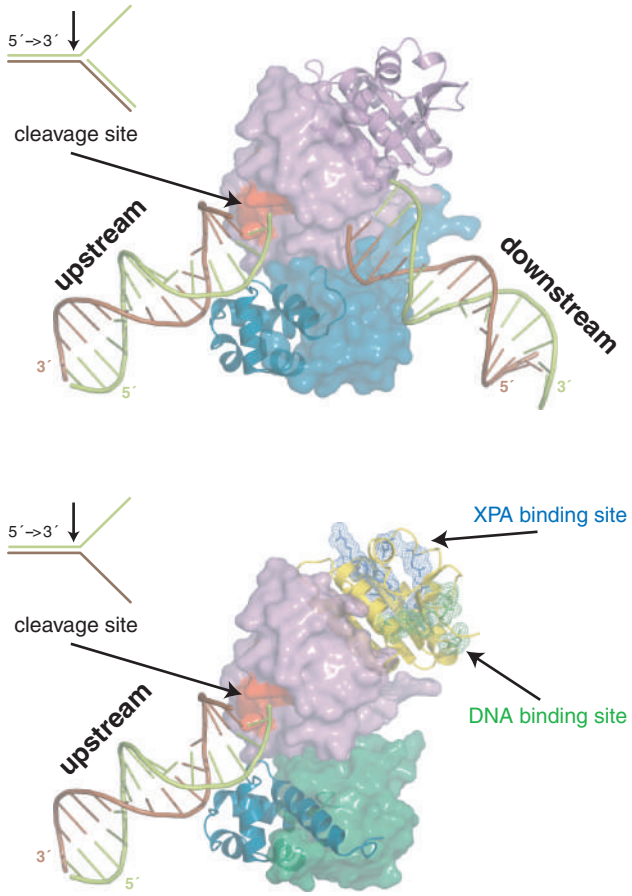


Figure 6. Models for the function of the archaeal homodimeric XPF (top) and human ERCC1/XPF heterodimer (bottom) constructed from the structure of the XPF homodimer bound to dsDNA (2bgw), the free structure of ERCC1/XPF C-terminal interacting domains (1z00) and the current structure of the ERCC1 central domain (2jpd). For the homodimer, one protomer is shown in a cartoon and the other in a surface representation. Accordingly, for the heterodimer ERCC1 is in a cartoon and XPF in a surface representation. The protein domains in both cases are colored as in Figure 1A.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Koos Jaspers and Jan Hoeijmakers for providing the full-length vectors of ERCC1 and XPF, and for many fruitful discussions. This work was financially supported by the Netherlands Organization for Scientific Research, Chemistry Council (NWO-CW), by the Center of Biomedical Genetics, the Netherlands, and by the EU project Spine2-Complexes (Contract no 031220). Funding to pay the Open Access publication charges for this article was provided by NWO-CW.

Conflict of interest statement. None declared.

REFERENCES

1. Nishino,T. and Morikawa,K. (2002) Structure and function of nucleases in DNA repair: shape, grip and blade of the DNA scissors. *Oncogene*, **21**, 9022–9032.
2. Ciccia,A., Ling,C., Coulthard,R., Yan,Z., Xue,Y., Meetei,A.R., Laghmani el,H., Joenje,H., McDonald,N. *et al.* (2007) Identification of FAAP24, a Fanconi anemia core complex protein that interacts with FANCM. *Mol. Cell*, **25**, 331–343.
3. Nishino,T., Komori,K., Ishino,Y. and Morikawa,K. (2003) X-ray and biochemical anatomy of an archaeal XPF/Rad1/Mus81 family nuclease: similarity between its endonuclease domain and restriction enzymes. *Structure (Camb.)*, **11**, 445–457.
4. Heyer,W.D., Ehmsen,K.T. and Solinger,J.A. (2003) Holliday junctions in the eukaryotic nucleus: resolution in sight? *Trends Biochem. Sci.*, **28**, 548–557.
5. Enzlin,J.H. and Scharer,O.D. (2002) The active site of the DNA repair endonuclease XPF-ERCC1 forms a highly conserved nuclease motif. *EMBO J.*, **21**, 2045–2053.
6. Niedernhofer,L.J., Odijk,H., Budzowska,M., van Drunen,E., Maas,A., Theil,A.F., de Wit,J., Jaspers,N.G.J., Beverloo,H.B. *et al.* (2004) The structure-specific endonuclease Ercc1-Xpf is required to resolve DNA interstrand cross-link-induced double-strand breaks. *Mol. Cell. Biol.*, **24**, 5776–5787.
7. Zhu,X.D., Niedernhofer,L., Kuster,B., Mann,M., Hoeijmakers,J.H.J. and de Lange,T. (2003) ERCC1/XPF removes the 3′ overhang from uncapped telomeres and represses formation of telomeric DNA-containing double minute chromosomes. *Mol. Cell*, **12**, 1489–1498.
8. Jaspers,N.G.J., Raams,A., Silengo,M.C., Wijgers,N., Niedernhofer,L.J., Robinson,A.R., Giglia-Mari,G., Hoogstraten,D. and Kleijer,W.J. (2007) First reported patient with human ERCC1 deficiency has cerebro-oculo-facio-skeletal syndrome with a mild defect in nucleotide excision repair and severe developmental failure. *Am. J. Hum. Genet.*, **80**, 457–466.
9. Niedernhofer,L.J., Garinis,G.A., Raams,A., Lalai,A.S., Robinson,A.R., Appeldoorn,E., Odijk,H., Oostendorp,R., Ahmad,A. *et al.* (2006) A new progeroid syndrome reveals that genotoxic stress suppresses the somatotroph axis. *Nature*, **444**, 1038–1043.
10. Newman,M., Murray-Rust,J., Lally,J., Rudolf,J., Fadden,A., Knowles,P.P., White,M.F. and McDonald,N.Q. (2005) Structure of an XPF endonuclease with and without DNA suggests a model for substrate recognition. *EMBO J.*, **24**, 895–905.
11. Nishino,T., Komori,K., Ishino,Y. and Morikawa,K. (2005) Structural and functional analyses of an archaeal XPF/Rad1/Mus81 nuclease: asymmetric DNA binding and cleavage mechanisms. *Structure*, **13**, 1183–1192.
12. Tsodikov,O.V., Enzlin,J.H., Scharer,O.D. and Ellenberger,T. (2005) Crystal structure and DNA binding functions of ERCC1, a subunit of the DNA structure-specific endonuclease XPF-ERCC1. *Proc. Natl Acad. Sci. USA*, **102**, 11236–11241.
13. de Laat,W.L., Appeldoorn,E., Jaspers,N.G.J. and Hoeijmakers,J.H.J. (1998) DNA structural elements required for ERCC1-XPF endonuclease activity. *J. Biol. Chem.*, **273**, 7835–7842.
14. Tripsianes,K., Folkers,G.E., AB,E., Das,D., Odijk,H., Jaspers,N.G.J., Hoeijmakers,J.H.J., Kaptein,R. and Boelens,R. (2005) The structure of the human ERCC1/XPF interaction domains reveals a complementary role for the two proteins in nucleotide excision repair. *Structure*, **13**, 1849–1858.
15. Gillet,L.C. and Scharer,O.D. (2006) Molecular mechanisms of mammalian global genome nucleotide excision repair. *Chem. Rev.*, **106**, 253–276.
16. Missura,M., Buterin,T., Hindges,R., Hubscher,U., Kasparkova,J., Brabec,V. and Naegeli,H. (2001) Double-check probing of DNA bending and unwinding by XPA-RPA: an architectural function in DNA repair. *EMBO J.*, **20**, 3554–3564.
17. Riedl,T., Hanaoka,F. and Egly,J.M. (2003) The comings and goings of nucleotide excision repair factors on damaged DNA. *EMBO J.*, **22**, 5293–5303.
18. Volker,M., Mone,M.J., Karmakar,P., van Hoffen,A., Schul,W., Vermeulen,W., Hoeijmakers,J.H.J., van Driel,R., van Zeeland,A.A. *et al.* (2001) Sequential assembly of the nucleotide excision repair factors in vivo. *Mol. Cell*, **8**, 213–224.
19. Li,L., Peterson,C.A., Lu,X. and Legerski,R.J. (1995) Mutations in XPA that prevent association with ERCC1 are defective in nucleotide excision repair. *Mol. Cell. Biol.*, **15**, 1993–1998.
20. de Jong,R.N., Daniels,M.A., Kaptein,R. and Folkers,G.E. (2007) Enzyme free cloning for high throughput gene cloning and expression. *J. Struct. Funct. Genomics* [Epub ahead of print].
21. Folkers,G.E., van Buuren,B.N. and Kaptein,R. (2004) Expression screening, protein purification and NMR analysis of human protein domains for structural genomics. *J. Struct. Funct. Genomics*, **5**, 119–131.
22. Delaglio,F., Grzesiek,S., Vuister,G.W., Zhu,G., Pfeifer,J. and Bax,A. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR*, **6**, 277–293.
23. Goddard,T.D. and Kneller,D.G. (2001) *SPARKY 3* University of California, San Francisco.
24. Cavanagh,J., Fairbrother,J.W., Palmer,G.A. III and Skelton,J.N. (1996) *Protein NMR Spectroscopy*. Academic Press, San Diego, CA, USA.
25. Herrmann,T., Güntert,P. and Wüthrich,K. (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.*, **319**, 209–227.
26. Cornilescu,G., Delaglio,F. and Bax,A. (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR*, **13**, 289–302.
27. Nederveen,A.J., Doreleijers,J.F., Vranken,W., Miller,Z., Spronk,C.A.E.M., Nabuurs,S.B., Güntert,P., Livny,M., Markley,J.L. *et al.* (2005) RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. *Proteins*, **59**, 662–762.
28. Brünger,A.T., Adams,P.D., Clore,G.M., DeLano,W.L., Gros,P., Grosse-Kunstleve,R.W., Jiang,J.S., Kuszewski,J., Nilges,M. *et al.* (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta. Crystallogr. D Biol. Crystallogr.*, **54** (Pt 5), 905–921.
29. DeLano,W. (2002) *The PyMOL Molecular Graphics System*. DeLano Scientific, San Carlos, CA, USA.
30. Singh,S., Folkers,G.E., Bonvin,A.M.J.J., Boelens,R., Wechselberger,R., Niztayev,A. and Kaptein,R. (2002) Solution structure and DNA-binding properties of the C-terminal domain of UvrC from E.coli. *EMBO J.*, **21**, 6257–6266.
31. Jonker,H.R., Wechselberger,R.W., Boelens,R., Folkers,G.E. and Kaptein,R. (2005) Structural properties of the promiscuous VP16 activation domain. *Biochemistry*, **44**, 827–839.
32. Sijbers,A.M., de Laat,W.L., Ariza,R.R., Biggerstaff,M., Wei,Y.F., Moggs,J.G., Carter,K.C., Shell,B.K., Evans,E. *et al.* (1996) Xeroderma pigmentosum group F caused by a defect in a structure-specific DNA repair endonuclease. *Cell*, **86**, 811–822.
33. Gaillard,P.H. and Wood,R.D. (2001) Activity of individual ERCC1 and XPF subunits in DNA nucleotide excision repair. *Nucleic Acids Res.*, **29**, 872–879.
34. Li,L., Elledge,S.J., Peterson,C.A., Bales,E.S. and Legerski,R.J. (1994) Specific association between the human DNA repair proteins XPA and ERCC1. *Proc. Natl Acad. Sci. USA*, **91**, 5012–5016.
35. Buchko,G.W., Isern,N.G., Spicer,L.D. and Kennedy,M.A. (2001) Human nucleotide excision repair protein XPA: NMR spectroscopic studies of an XPA fragment containing the ERCC1-binding region and the minimal DNA-binding domain (M59-F219). *Mutat. Res.*, **486**, 1–10.
36. Force,A., Lynch,M., Pickett,F.B., Amores,A., Yan,Y.L. and Postlethwait,J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151**, 1531–1545.
37. Lynch,M. and Conery,J.S. (2003) The evolutionary demography of duplicate genes. *J. Struct. Funct. Genomics*, **3**, 35–44.
38. Roberts,J.A. and White,M.F. (2005) An archaeal endonuclease displays key properties of both eukaryal XPF-ERCC1 and Mus81. *J. Biol. Chem.*, **280**, 5924–5928.
39. Baker,N.A., Sept,D., Joseph,S., Holst,M.J. and McCammon,J.A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl Acad. Sci. USA*, **98**, 10037–10041.

THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 285, NO. 37, pp. 28893–28901, September 10, 2010
© 2010 by The American Society for Biochemistry and Molecular Biology, Inc. Printed in the U.S.A.

Structural Basis for Homodimerization of the Src-associated during Mitosis, 68-kDa Protein (Sam68) Qua1 Domain*[S]◆

Received for publication, March 23, 2010, and in revised form, May 19, 2010 Published, JBC Papers in Press, July 6, 2010, DOI 10.1074/jbc.M110.126185

N. Helge Meyer[‡§], Konstantinos Tripsianes[‡§], Michelle Vincendeau[¶], Tobias Madl[‡§], Fatiha Kateb[‡§], Ruth Brack-Werner[¶], and Michael Sattler[‡§1]

From the [‡]Institute of Structural Biology, Helmholtz Zentrum München, Ingolstädter Landstrasse 1, 85764 Neuherberg, the [§]Munich Center for Integrated Protein Science (CiPS[M]) at Department Chemie, Technische Universität München, Lichtenbergstrasse 4, 85747 Garching, and the [¶]Institute of Virology, Helmholtz Zentrum München, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany

Sam68 (Src-associated during mitosis, 68 kDa) is a prototypical member of the STAR (signal transducer and activator of RNA) family of RNA-binding proteins. STAR proteins bind mRNA targets and modulate cellular processes such as cell cycle regulation and tissue development in response to extracellular signals. Sam68 has been shown to modulate alternative splicing of the pre-mRNAs of CD44 and Bcl-xL, which are linked to tumor progression and apoptosis. Sam68 and other STAR proteins recognize bipartite RNA sequences and are thought to function as homodimers. However, the structural and functional roles of the self-association are not known. Here, we present the solution structure of the Sam68 Qua1 homodimerization domain. Each monomer consists of two antiparallel α-helices connected by a short loop. The two subunits are arranged perpendicular to each other in an unusual four-helix topology. Mutational analysis of Sam68 *in vitro* and in a cell-based assay revealed that the Qua1 domain and residues within the dimerization interface are essential for alternative splicing of a CD44 minigene. Together, our results indicate that the Qua1 homodimerization domain is required for regulation of alternative splicing by Sam68.

Sam68 [2] (Src-associated during mitosis, 68 kDa) (1) belongs to the STAR (signal transducer and activator of RNA) family of RNA-binding proteins, which also includes Qk1 (quaking 1), SF1 (splicing factor 1), and Gld-1 (germline development defective-1) (2). STAR family proteins link signaling pathways and many aspects of RNA metabolism (splicing, localization, and translation). They are regulated by post-translational modifications such as phosphorylation, acetylation, and arginine methylation (2).

Sam68 acts in post-transcriptional regulation of pre-mRNA splicing in response to extracellular signals (3). It is involved in a variety of pathways, including insulin and T-cell receptor signaling (4), and plays a key role in cell cycle regulation (5). Sam68 exhibits binding specificity for homopolymeric poly(U) RNA and has been shown to recognize UAAA or UUUA sequences with high affinity as determined by Systematic Evolution of Ligands by EXponential Enrichment (SELEX) and *in vivo* cross-linking (6, 7). Post-translational modifications can regulate Sam68 function by critically affecting the accessibility to RNA (8, 9). Tyrosine phosphorylation by Src kinase during mitosis enhances the interaction of Sam68 with Ras-GAP (10) but prevents its association with RNA. On the other hand, acetylation of lysine residues by histone acetyltransferases enhances RNA binding (11). Finally, overexpression of Sam68 has been linked to prostate cancer, cell proliferation, and survival (12).

Sam68 has been identified as a key determinant in the alternative splicing of various important RNA targets, like CD44 (13) and Bcl-x_L (14), which are linked to apoptosis and cancer. In particular, alternative splicing of CD44 impacts embryonic development and immune response (15–18). Up to 10 variant exon sequences can be included in the mature CD44 mRNA. Among them, variable exon 5 (exon v5) inclusion is associated with tumor progression and T-cell activation (17, 19). Sam68 is a target for phosphorylation by extracellular-signal-regulated kinase (ERK), which promotes inclusion of CD44 exon v5 in response to Ras activation by phorbol ester stimulation. Sam68 binds two RNA sequences within exon v5 and the preceding intron, respectively (9, 13). Recent studies indicate that Sam68 interacts with the heterodimeric splicing factor U2AF (U2 small nuclear ribonucleoprotein particle auxiliary factor) (9). Recognition of the 3′ splice site by U2AF is a key step in spliceosome formation. The interaction with Sam68 is thought to stabilize U2AF binding to its cognate RNA elements. Interestingly, phosphorylation of Sam68 by ERK interferes with RNA binding and reduces pre-mRNA occupancy by U2AF. Altogether, these data indicate that Sam68 can regulate alternative splicing in a signal-dependent manner.

Sam68 shares high similarity with SF1, another binding partner of U2AF. The so-called STAR domain, also referred to as GRP33/SAM68/GLD1 domain, consists of an hnRNP K homology (KH) domain flanked by two domains N- and C-terminal of

* This work was supported by a grant from the Deutsche Forschungsgemeinschaft and grants from the Alexander von Humboldt foundation (to K. T. and F. K.) and an EMBO fellowship (to T. M.).

◆ This article was selected as a Paper of the Week.

The atomic coordinates and structure factors (code 2XA6) have been deposited in the Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ (http://www.rcsb.org/).

The chemical shifts can be accessed through the Biological Magnetic Resonance Bank (BMRB) under BMRB accession number 16969.

[S] The on-line version of this article (available at http://www.jbc.org) contains supplemental Figs. 1–7.

[1] To whom correspondence should be addressed. Tel.: 49-89-289-13418; Fax: 49-89-289-13869; E-mail: sattler@helmholtz-muenchen.de.

[2] The abbreviations used are: Sam68, Src-associated during mitosis, 68-kDa protein; STAR, signal transducer and activator of RNA; SF1, splicing factor 1; Gld-1, germline development defective-1; U2AF, U2 small nuclear ribonucleoprotein particle auxiliary factor; KH, hnRNP K homology; HSQC, heteronuclear single-quantum correlation; r.m.s.d., root mean square deviation.

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

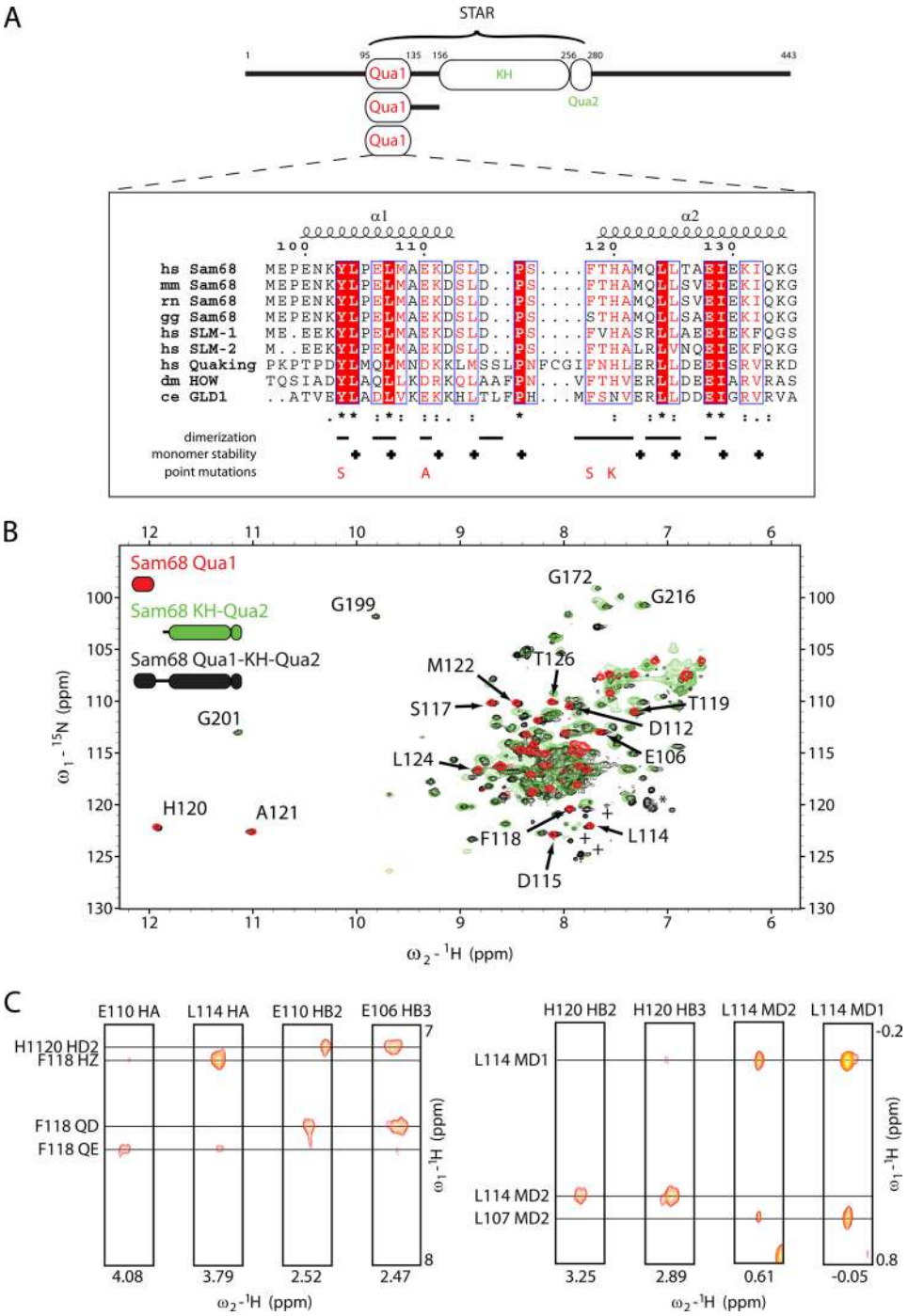


FIGURE 1. **Topology and NMR spectra of the Sam68 STAR domain.** *A*, domain organization of Sam68. The Qua1 proteins used for NMR are indicated. The sequence alignment of the Qua1 domains of different members in the STAR family is shown, and residue numbers are given for Sam68 Qua1. The secondary structure of Sam68 Qua1 is depicted above the alignment. Residues important for monomer and dimer stability are indicated below, by − and +, respectively. Conserved residues that were mutated are indicated. Sequences were aligned using the program ClustalW (51) and analyzed using the program ESpript (52). Abbreviations used are: *hs*, *H. sapiens*; mm, *Mus musculus*; *rn*, *Rattus norvegicus*; *gg*, *Gallus gallus domesticus*; *dm*, *Drosophila melanogaster*; *ce*, *Caenorhabditis elegans*. *B*, overlay of a $^{1}H,^{15}N$ TROSY spectrum of perdeuterated Qua1-KH-Qua2 (residues 95–280) (*blue*) and $^{1}H,^{15}N$ HSQC spectra of Qua1 (residues 95–135) (*red*) and KH-Qua2 (residues 147–280) (*green*). Note that the linker connecting the Qua1 and KH domains (135–156) is not included in the Qua1 protein. Signals arising from arginine side chains are marked with an *asterisk*, and some small peaks of degradation products are marked with +. *C*, intermolecular NOEs recorded on an asymmetrically $^{15}N/^{13}C$-labeled sample.

the KH domain, referred to as Qua1 and Qua2, respectively (2, 3). Interestingly, SF1 is the only known member devoid of the Qua1 subdomain and functions as a monomer (20). The KH domain is one of the major RNA binding motifs in eukaryotic cells (21). The structural basis for the recognition of single-stranded RNA by the KH-Qua2 domain of SF1 has been described, indicating that the Qua2 domain extends the KH fold and that the KH-Qua2 tandem domain is essential for sequence-specific RNA recognition (22). The solution structure of the free form of the KH-Qua2 region of Quaking showed that the Qua2 helix does not contact the KH domain in the absence of RNA (23). Biochemical data and *in vivo* studies suggest that the Qua1 domain can oligomerize (24–27). However, the structural basis for oligomerization and the role for the *in vivo* function of STAR proteins are unknown.

Here, we present the NMR solution structure of the Sam68 Qua1 domain. Qua1 forms a homodimer composed of a perpendicular interaction of two helical hairpins. A network of hydrophobic and electrostatic interactions stabilizes the dimer interface. Based on NMR and biophysical data, we show that the Qua1 domain is sufficient for homodimerization of Sam68 *in vitro*. Cell-based splicing assays identify critical residues in the dimer interface and reveal that Qua1 is necessary for the function of Sam68 in alternative splicing.

EXPERIMENTAL PROCEDURES

Sample Preparation—Homo sapiens Sam68 Qua1(95–135), Qua1(95–156), KH-Qua2(147–280), and Qua1-KH-Qua2(95–280), which were derived from pcDNA 3.1 HsSam68 described in Ref. 13, were cloned in a modified pETM-11 vector (European Molecular Biology Laboratory) containing an additional N-terminal double Z-tag and expressed in *Escherichia coli* BL21(DE3). Point mutations were introduced in pcDNA 3.1 HsSam68 and pETM-11 ZZ Qua1(95–135) using the QuikChange site-directed mutagenesis (Invitrogen) protocol. The corresponding proteins contain two additional residues at the N-terminal tobacco etch virus protease cleavage site, which are derived from the expression vector.

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

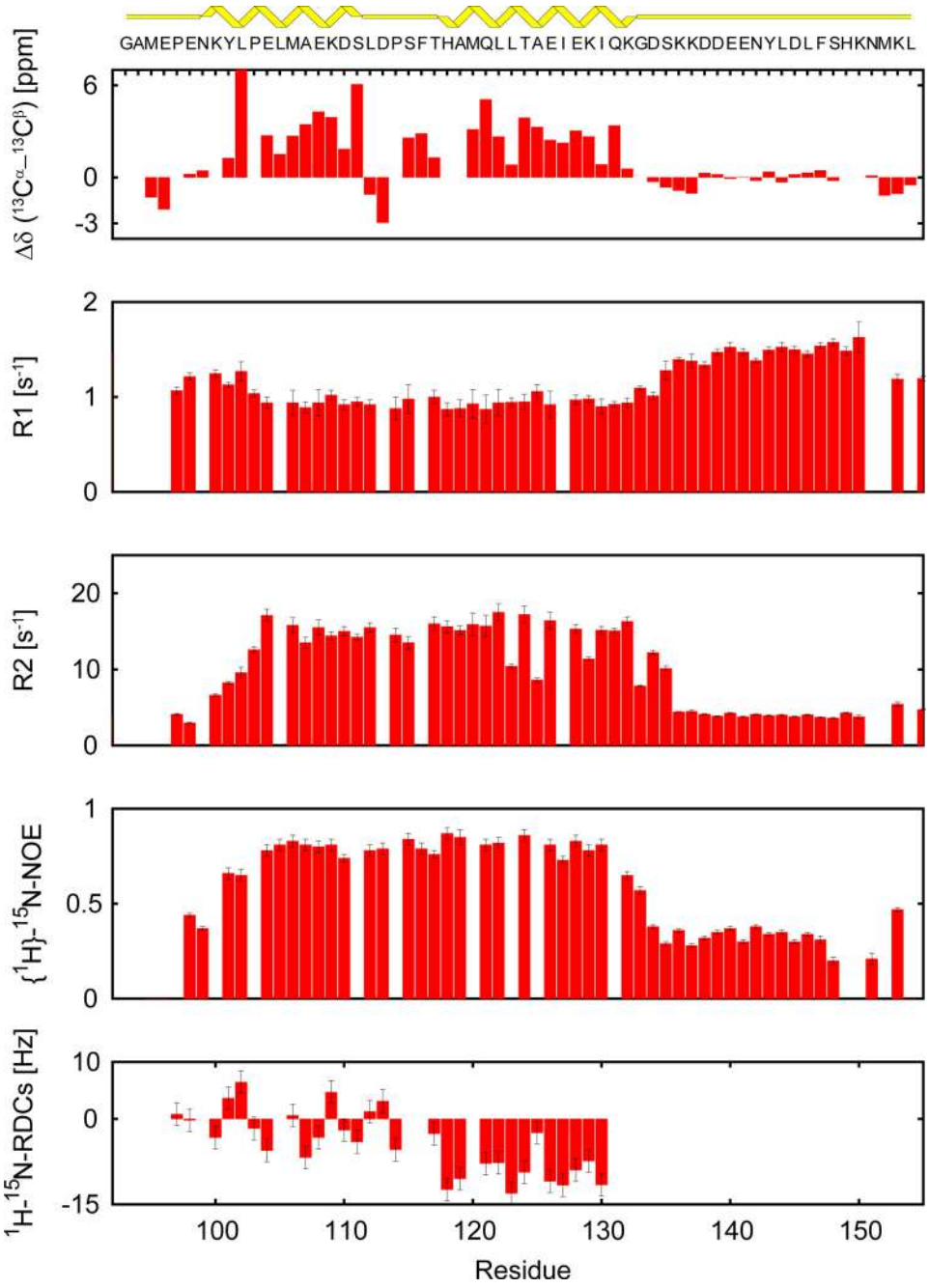


FIGURE 2. **NMR characterization of the Sam68 Qua1 domain.** ^{13}C secondary chemical shifts, ^{15}N R1, R2 relaxation rates, {^{1}H}-^{15}N heteronuclear NOE, and ^{1}H-^{15}N residual dipolar couplings (*RDCs*) are plotted *versus* Sam68 Qua1 residue numbers. The secondary structure is indicated at the *top*. *Error bars* for ^{15}N relaxation data are estimated from the noise level as described in Ref. 33. *Error bars* for RDCs are estimated based on spectral resolution.

Bacteria were grown in LB medium for preparation of unlabeled sample or M9 medium supplemented with ^{13}C-labeled glucose and/or [^{15}N]H_4Cl for uniformly labeled samples, respectively. The protein was expressed at 20 °C for 16 h after induction with 250 mM IPTG.

Proteins were purified using Ni^{2+} affinity chromatography (nickel-nitrilotriacetic acid, Qiagen). After cleavage of the His and the Z-tag with tobacco etch virus protease, the tags were separated from the protein by a second Ni^{2+} affinity step. Qua1 samples were heat-shocked for 5 min at 85 °C. The samples were further purified by size exclusion chromatography on a HiLoad 16/60 Superdex 75 (GE Healthcare) and kept in 10 mM phosphate, pH 6.5, 100 mM NaCl.

An asymmetrically labeled sample to record intermolecular NOEs was prepared as follows. Equimolar amounts of unlabeled and ^{15}N/^{13}C uniformly labeled Qua1 (residues 95–135) were mixed and incubated with 1% SDS at 85 °C for 10 min. After slowly cooling down, SDS was removed by dilution and ultra filtration using a 15-ml Amicon (Millipore) with a molecular weight cutoff of 5,000.

NMR Spectroscopy—NMR measurements were carried out at 298 K on a Bruker Avance III 750-MHz spectrometer equipped with a TXI probe head, a 600-MHz spectrometer equipped with a TCI cryo-probe head or on an Avance 900 instrument equipped with a TXI cryo-probe head. Spectra were processed with NMRPipe (28) and analyzed with Sparky (29). Backbone assignment was done semiautomatically using MARS (30). For backbone and side chain assignment CBCA(CO)NH, CBCANH and (H)CCH-TOCSY spectra were recorded (31). Distance information was obtained from ^{15}N- and ^{13}C-edited NOESY spectra with a mixing time of 70 ms. To distinguish inter- and intramolecular NOEs, a set of isotope-edited and filtered NOESY spectra was recorded (31, 32). Experiments were carried out using a 1 mM uniformly ^{15}N/^{13}C-labeled or 2 mM asymmetrically ^{15}N/^{13}C-labeled Qua1(95–156) sample, respectively. ^{15}N R_1 and R_2 relaxations rates and {^{1}H}-^{15}N heteronuclear NOE data were measured at a 750-MHz proton Larmor frequency and 298 K as described (33). H^{α}-C^{α}, N-C′ and H^{N}-N residual dipolar couplings were recorded using HNCO-based NMR experiments (34) with a 1 mM Qua1(95–135) sample that was aligned in a medium containing 15 mg/ml Pf1 phage (Profos AG, Regensburg, Germany) as described (35). Paramagnetic relaxation enhancements were measured from saturation recovery ^{1}H,^{15}N HSQC experiments (recovery times between 0.01 and 4.0 s) at concentrations of 0, 1, 2, 3, 4, 5, 7, and 10 mM of the soluble paramagnetic agent gadolinium diethylenetriaminepentaacetic acid bismethylamide (Gd(DTPA-BMA)). Back-calculation and data analysis were carried out according to Ref. 36.

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

Structure Calculation—Automated NOE cross-peak assignment was performed using the software CYANA 3.0 (37). Automatically assigned NOEs and completeness of the NOE cross-peaks were manually inspected. Homodimer symmetry is explicitly taken into account for network anchoring, and identical conformation of the two monomers is ensured by dihedral angle difference restraints for all corresponding torsion angles. Additionally, a symmetric relative orientation of the two monomers is maintained by distance difference restraints between symmetry-related intermolecular C^{α}-C^{α} distances (38). Distance restraints from the CYANA calculation- and TALOS+ (39)-derived torsion angles and the residual dipolar coupling restraints were used in a water refinement calculation (40) with Aria 2.2 (41). The quality of the structure ensemble was validated using the iCING[3] web server as well as PROCHECK (43) and WHATCHECK (44). Molecular images were generated using PyMOL (45).

Circular Dichroism (CD) Spectroscopy—Temperature series of far-UV (190–250 nm) CD spectra were recorded on 100 μM Qua1 wild type or mutant proteins in 20 mM phosphate, pH 6.5, 50 mM NaCl using a JASCO J-715 spectropolarimeter.

Cell Culture and Transfections—U138MG cells were cultured as described previously (46). All overexpression experiments were performed in 6-well plates. Cells were seeded at a density of 2×10^5 cells/well 1 day prior to transfection and cultured for 30 h after transfection. Transfections were performed using FuGENE® HD transfection reagent (Roche Diagnostics) according to the manufacturer's instructions. For co-transfection experiments, 2 μg of CD44 v5 minigene-plasmid DNA (13) was co-transfected with 300 ng of plasmids expressing Sam68 or wild type or mutants. As a control, the CD44 v5 minigene-plasmid DNA was co-transfected with 300 ng of the pcDNA 3.1 expression plasmid. Expression of all proteins was checked by Western blot analysis (data not shown).

RT-PCR Analysis—Cytoplasmic RNA was prepared using the PARIS kit (Ambion) according to the manufacturer's protocol. 1 units/μg of RNA was transcribed with SuperScript II (Invitrogen) according to the manufacturer's protocol, using random hexamers. RT-PCR analyses were carried out as described in Ref. 13 using 25 PCR cycles. RT-PCR bands were quantified densitometrically using the ImageJ Software. Three independent experiments were performed.

RESULTS

Qua1 Is Sufficient for Sam68 Homodimerization—It has been shown that the STAR domain of Sam68 (Fig. 1*A*) and other members of the STAR family dimerize (24–27). The apparent molecular mass estimated from size exclusion chromatography of ~40 kDa is consistent with a dimer of the Sam68 STAR domain (not shown). To identify the region that mediates dimerization, we expressed proteins comprising the Qua1, KH-Qua2, and Qua1-KH-Qua2 domains in *E. coli*. An overlay of ^{1}H,^{15}N correlation NMR spectra of the Qua1 (residues 95–135) and KH-Qua2 (residues 147–280) subdomains with that of the STAR domain (residues 95–280) shows that chemical shifts seen in the smaller domains are very similar to those of common residues in the STAR domain. This indicates that there are no strong contacts between Qua1 and KH-Qua2 and that these two regions are structurally autonomous (Fig. 1*B*).

TABLE 1
Structural statistics of the Sam68 Qua1 homodimer

Statistics are given for the 20 lowest energy structures after water refinement out of 100 calculated. The CNS E_{repel} function was used to simulate van der Waals interactions with an energy constant of 25 kcal mol^{-1} Å^{-4} using PROLSQ van der Waals radii (40). r.m.s.d. and PROCHECK values apply for residues 101–135.

NOE-based distance restraints[a]	
Intraresidual, sequential	950
Medium range ($2 \leq \|i-j\| \leq 4$)	504
Long range ($\|i-j\| \geq 5$)	316
Intermolecular	456
Total	2226
Other restraints	
$\phi+\psi$ dihedral angle restraints	128
Residual dipolar coupling restraints (H^N-N, N-C′, H^{α}-C^{α})	128
Coordinate precision r.m.s.d.	
Backbone (Å)	0.32 ± 0.09
Heavy atom (Å)	0.97 ± 0.16
Consistency (structure vs. restraints)	
r.m.s.d. (Å) from experimental distance restraints[a]	0.019 ± 0.002
r.m.s.d. (°) from experimental torsion angle restraints[b]	0.911 ± 0.104
RDC Q-factor[c]	0.230 ± 0.005
	Structure Z-score
WHATCHECK[d]	
First generation packing quality	0.497
Second generation packing quality	−0.028
Ramachandran plot appearance	0.010
χ_1/χ_2 rotamer normality	−2.057
Backbone conformation	1.325
Ramchandran plot[d]	
Most favored regions	96.0%
Allowed regions	4.0%
Generously allowed regions	0.0%
Disallowed regions	0.0%

[a] Distance restraints were employed with a soft square well potential using an energy constant of 50 kcal mol^{-1} Å^{-2}. No distance restraint was violated by more than 0.5 Å.
[b] Torsion angle restraints derived from TALOS (39) were applied to φ, ψ backbone angles using energy constants of 200 kcal mol^{-1} radians^{-2}. No dihedral angle restraint was violated by more than 5°.
[c] Residual dipolar couplings (RDCs) were employed with a harmonic potential using an energy constant of 0.5 kcal mol^{-1} Hz^{-2}. Q-factor as defined in Ref. 42.
[d] PROCHECK (43) and WHATCHECK (44) were used to determine the quality of the structure. Positive WHATCHECK Z-scores indicate that structure is better than average.

The ^{1}H NMR frequencies in the ^{1}H,^{15}N HSQC spectrum of Qua1 (residues 95–156) are mainly found within a small region centered around 8 ppm, indicating the presence of mostly helical and random coil conformation. Analysis of the $^{13}C^{\alpha/\beta}$ secondary chemical shifts reveals that residues 95–135 comprise two α-helices that are interconnected by a short loop. The 20 C-terminal amino acids (residues 136–156) do not exhibit any secondary structure. ^{15}N R_1 and R_2 relaxation rates as well as {^{1}H}-^{15}N heteronuclear NOE data show that the C terminus is highly flexible, whereas the loop connecting the two helices has a slightly increased flexibility when compared with the helical segments (Fig. 2). Thus, only residues 95–135 define the globular fold of the Qua1 domain. A tumbling correlation time of $\tau_c = 9.5$ ns was estimated from the ^{15}N R_2/R_1 relaxation rates ratio for Qua1, consistent with the molecular mass expected for a homodimer (14 kDa). Even at concentrations of 10 μM, there is no significant change in the ^{15}N relaxation rates and thus τ_c (data not shown), indicating that the dimerization constant is at least in the low micromolar to nanomolar range. Additionally,

[3] G. W. Vuister, J. F. Doreleijers, and A. W. Sousa da Silva, manuscript in preparation.

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

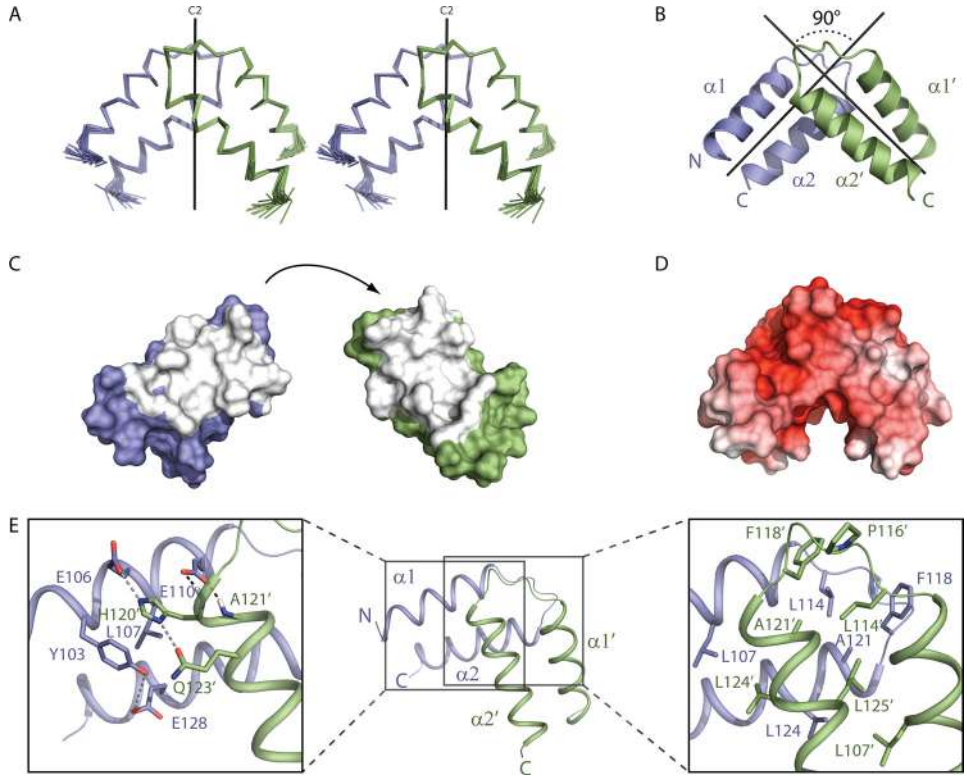


FIGURE 3. **NMR structure of the Sam68 Qua1 homodimer.** *A*, stereo view of the ensemble of the 20 lowest energy structures. *B*, ribbon representation of the Sam68 Qua1 dimerization domain. *C*, surface representation of two separated Qua1 monomers, showing a view onto the dimer interface (*white*). *D*, surface representation colored according to electrostatic surface potential at ±7 k_B T/e (temperature/electron charge) for positive (*blue*) or negative (*red*) charge potential using the program APBS (53). *E*, close-up view of the interface. Side chains of key residues in the dimer interface are shown as *sticks*. Hydrogen bonds between the Glu-110 side chain and the His-120 and Ala-121 amides are indicated by *black dashed lines*. Potential hydrogen bonds and electrostatic interactions are shown by *gray dashed lines*.

the apparent molecular weight determined by size exclusion chromatography matches well with a homodimer (supplemental Fig. 1). These observations suggest that the Qua1 domain is sufficient for homodimerization.

For structural analysis and to further confirm that Qua1 is a homodimer, we attempted to identify intermolecular NOEs. In the case of a symmetric homodimer, the NOESY spectrum contains both intramolecular and intermolecular cross-peaks. To unambiguously distinguish intermolecular from intramolecular NOEs, isotope-filtered experiments were recorded on an asymmetrically $^{15}N/^{13}C$-labeled sample (31, 32). To obtain such a sample, equimolar amounts of uniformly $^{15}N/^{13}C$-labeled and unlabeled Qua1 are mixed. Random exchange of the subunits yields 50% of asymmetrically isotope-labeled dimers. Isotope-edited/filtered NMR experiments will detect exclusively intermolecular NOEs for this population, whereas no signals are observed for symmetrically labeled or unlabeled homodimers or monomeric species. Thermal denaturation in presence of 1% SDS and refolding of equal amounts of labeled and unlabeled Qua1 resulted in a fingerprint spectrum identical to that of the native protein, indicating that the fold was restored (data not shown). After refolding, more than 100 intermolecular cross-peaks were observed in an isotope-edited/filtered NOESY experiment (Fig. 1*C*). Taken together, these data establish that Qua1 is a stable homodimer.

Structure of the Sam68 Qua1 Homodimer—We determined the solution structure of the Sam68 Qua1 homodimer (comprising residues 95–135) by heteronuclear triple-resonance NMR spectroscopy using uniformly as well as asymmetrically labeled $^{15}N/^{13}C$-labeled protein. The structure is defined based on 2226 experimental distance restraints, including 2 × 228 intermolecular distance restraints, derived from ^{13}C- and ^{15}N-edited NOESY HSQC spectra as well as $^{13}C/^{15}N$-filtered, ^{15}N- or ^{13}C-edited NOESY-HSQC spectra. A summary of the structural and restraint statistics is given in Table 1. The ensemble of the 20 lowest energy structures obtained after water refinement is shown in Fig. 3*A*. The quality of the structures was further validated by comparison of measured and back-calculated relaxation rate enhancements upon the addition of the paramagnetic co-solvent Gd(DTPA-BMA) (36), which showed excellent agreement (supplemental Fig. 2).

Each Qua1 monomer comprises two α-helices that are aligned in an antiparallel fashion with a tilt angle of ~30°, connected by a short loop. Hydrophobic residues are spaced every 3–4 residues apart on each helix and interact with residues arrayed on the other helix of the same monomer, *i.e.* Leu-104, Leu-107, Lys-111 and Leu-114 in helix α1 contact Ile-132, Ile-129, Leu-125, and Met-122 in helix α2, respectively (supplemental Fig. 3). This arrangement is reminiscent of a short "zipper" where the extensive hydrophobic contacts lead to a tight packing of the two helices of each monomer. A potential side-chain hydrogen bond between the invariant Tyr-103 (helix α1) and Glu-128 (helix α2) further stabilizes the helical hairpin. The lack of the corresponding hydrogen bond may be consistent with a lethal phenotype of the E48G mutation in the mouse Qk1 paralog (Glu-48 corresponds to Sam68 Glu-128). Another important feature of this topology is the highly conserved Pro-116 in the loop connecting the two helices, which allows the reversal of the peptide backbone.

The Qua1 dimer is formed by perpendicular stacking of the two monomers with a C2 symmetry (Fig. 3, *A* and *B*). Numerous hydrophobic contacts, which mainly involve the loop region connecting the two helices, stabilize the dimer interface. Below this largely hydrophobic contact area, cross-subunit hydrogen bonds and electrostatic contacts between the N-terminal end of helix α2 and the α1′/α2′ helices of the other monomer provide additional intersubunit interactions (Fig. 3). The interface area covers 624

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

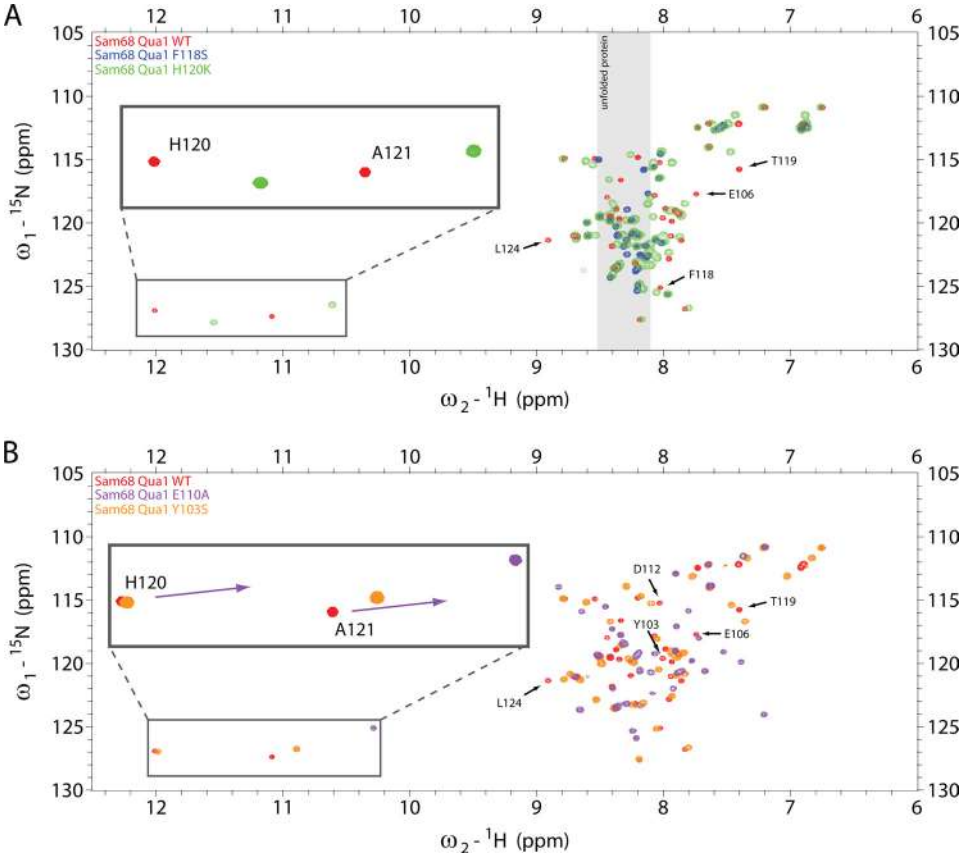


FIGURE 4. **NMR spectra of wild type and mutant Sam68 Qua1.** *A* and *B*, overlay of the ^{1}H,^{15}N HSQC spectra of the different Qua1 mutants that disrupt the overall fold (*A*) or that have mainly local structural effects (*B*). *A*, the F118S mutant is unfolded as shown by a drastically reduced chemical shift dispersion. The typical chemical shift range for unfolded proteins is indicated by a *gray box*. The H120K mutant shows two sets of NMR signals, which account for folded and unfolded moieties of the protein (see supplemental Fig. 4). *B*, the mutants Y103S and E110A are folded. However, the intermolecular hydrogen bond between the Glu-110 side chain and the backbone amides of His-120 and Ala-121 cannot form in the E110A mutant, resulting in an upfield shift of the backbone amide resonances for those residues (indicated by *arrows*). The E110A mutant may lead to a local rearrangement of the aromatic side chains of His-120 and indirectly of Tyr-103, which is reflected by the overall chemical shift differences when compared with the wild type protein. Similarly, the relatively large changes of the NMR chemical shifts in the Y103S mutant may be attributed to the removal of aromatic ring current effects.

Å^2/monomer, which accounts for more than 18% of the total surface area of one monomer (Fig. 3*C*).

A key determinant of the dimer formation is the almost invariant Phe-118 in the α1-α2 loop. Its aromatic side chain stacks with the Leu-114′ side chain from the other monomer (Fig. 3*E*, *right*). The side chains of Leu-107, Ala-121, Leu-124, and Leu-125 mediate additional hydrophobic interactions stabilizing the dimer interface. Near the intersection of the α2/α2′ helices, the His-120′ side chain contacts the aromatic ring of Tyr-103 in the other monomer. Apart from this hydrophobic network, cross-subunit hydrogen bonds between Glu-110 and the backbone amide protons of Ala-121′ and His-120′ are observed (Fig. 3*E*, *left*). This is reflected by large downfield shifts of the ^{1}H NMR frequencies of the backbone amides of Ala-121 and His-120 (Fig. 4). Furthermore, the side chain of His-120′ is in close proximity to the side chains of Glu-106 and Gln-123′, implying electrostatic stabilization (Fig. 3*E*, *left*). Charged residues are distributed at the periphery and at the opposite side of the interface, giving an overall negative charge to the Qua1 homodimer (Fig. 3*D*).

An arrangement of two antiparallel helices is a common feature in a variety of different proteins and protein folds. However, structural similarity searches (47) did not reveal any structure in the Protein Data Bank (PDB) database with a similar perpendicular arrangement of two helical hairpins, indicating that the Qua1 homodimer adopts a novel four-helix dimer topology.

Mutational Analysis of the Dimer Interface—We confirmed the dimer contacts seen in the structure by mutational analysis. Various single point mutations were designed, and ^{1}H,^{15}N HSQC experiments of the corresponding proteins were recorded to monitor the influence of the mutation on the overall fold. The capability to form homodimers was determined by analysis of apparent local correlation times, τ_c^{app}, estimated from the ^{15}N relaxation data. The τ_c^{app} values reflect the size of the protein tumbling in solution, and the distribution of τ_c^{app} therefore allows estimating the ratio between the monomeric and the dimeric forms (supplemental Fig. 4).

Mutation of Phe-118 (F118S) leads to a virtually complete loss of structure and dimerization (Fig. 4*A*). Remarkably, at concentrations above 1 mM, a small portion of the protein can still dimerize, as indicated by weak NMR signals in the ^{1}H,^{15}N HSQC, which reflect the folded dimer (data not shown). The H120K variant is in slow exchange between unfolded protein and structurally intact dimer, as indicated by two sets of signals with distinct ^{15}N relaxation properties that are observed in the ^{1}H,^{15}N HSQC spectra (Fig. 4*A*). The Y103S and E110A variants maintain the overall fold, and the protein still forms dimers (Fig. 4*B*). However the distribution of local correlation times, τ_c^{app}, derived from the NMR relaxation data (supplemental Fig. 4) is slightly decreased in both mutants when compared with the wild type, consistent with a fast exchange between monomeric and dimeric species. The effect of the E110A mutation on dimerization is more pronounced as the average τ_c^{app} measured for the E110A variant is lower than for the Y103S mutant. The E110A Qua1 domain cannot form intermolecular hydrogen bonds to the backbone amides of Ala-121′ and His-120′. This results in an upfield shift of the respective resonances seen in the ^{1}H,^{15}N HSQC spectrum of the E110A mutant (Fig. 4*B*). Thermal denaturation of the Qua1 mutants followed by CD spectroscopy demonstrates that the thermal stability of the Y103S and E110A variants is decreased when compared with

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

the wild type protein (supplemental Fig. 5), although both mutants are still mainly dimeric. In contrast, the F118S and H120K mutants are highly unstable or already unfolded at room temperature (data not shown).

In summary, the structure-guided mutagenesis shows a good correlation between the importance of contacts in the interface and dimer formation. Specifically, the F118S mutation completely disrupts the dimer, whereas the other mutations appear to have more local effects only destabilizing the Qua1 dimer.

Functional Qua1 Domain Is Required for Splicing Regulation by Sam68—Sam68 has previously been shown to increase the inclusion of the CD44 exon v5 in mRNAs produced from a CD44v5 minigene (Fig. 5*A*) (13). Here, we used this assay to investigate whether the Qua1 domain is required for regulation of splicing by Sam68. These experiments were performed in the human astrocytic cell line U138MG because human astrocytes were previously reported to express relatively low levels of endogenous Sam68 (48) (supplemental Fig. 6). Overexpression of wild type full-length Sam68 led to a more than 3-fold increase of exon v5 inclusion (Fig. 5, *B* and *C*). As expected, a mutant lacking the KH region, which mediates RNA binding, failed to increase exon v5 inclusion. Interestingly, the mutant lacking the Qua1 domain was similarly ineffective as the ΔKH mutant in this assay. This demonstrates that the Qua1 domain is essential for Sam68 activity *in vivo*. The same effect is observed for the F118S variant, which *in vitro*, in the context of the Qua1 domain, does not dimerize. H120K, which is disturbed in dimerization *in vitro* (supplemental Fig. 4), has significantly reduced splicing activity. In contrast, the E110A mutant is only slightly destabilized *in vitro*, and, consistently, supports alternative splicing almost as the wild type protein. Although mutation of Tyr-103 has the least effect on dimerization *in vitro*, it seems to be crucial for Sam68 activity as the Y103S variant is virtually inactive in the splicing assay.

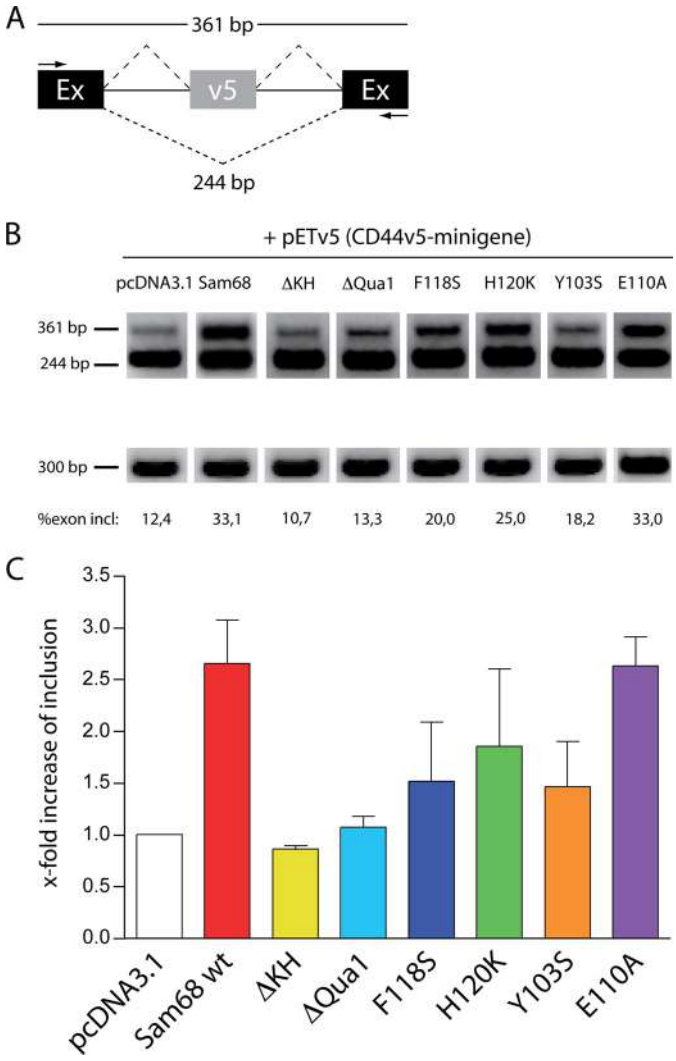


FIGURE 5. **Mutational analysis of the Sam68 Qua1 domain reduces CD44 exon v5 inclusion *in vivo*.** *A*, schematic presentation of the CD44 v5 minigene organization. *Black boxes* display the insulin exons (*Ex*); the *gray box* represents CD44 exon v5. PCR primer positions are indicated as *arrows*. Inclusion of CD44 v5 leads to a PCR product with a predicted size of 361 bp, whereas exclusion of exon v5 results in a 244-bp PCR product. *B*, RT-PCR analysis of minigene splicing patterns from cytoplasmic RNA of U138MG cells co-transfected with the CD44 v5 minigene construct pETv5 (13) and with plasmids for expression of wild type or mutant Sam68 proteins. Cells were co-transfected with the minigene construct and the empty vector plasmid (pcDNA3.1) as control. The expression levels of a housekeeping gene (RNA polymerase II) are shown for comparison. Values for exon inclusion (expressed as the percentage of variant exon-containing RT-PCR products relative to total RT-PCR products) are displayed. *C*, RT-PCR bands were quantified densitometrically using the ImageJ Software. Each Sam68 protein was assayed in three independent co-transfection experiments. The graph shows the *x*-fold increase of the percentage of exon inclusion for the RT-PCR products in U138MG cells, relative to the control. Columns represent the mean results of three independent experiments, and *error bars* represent the S.E.

DISCUSSION

We show that the Sam68 Qua1 domain adopts a unique four-helix dimer fold with the two monomers stacking perpendicularly. The dimer interface is stabilized by a combination of hydrophobic interactions and cross-subunit hydrogen bonds. Residues that mediate these interactions are highly conserved (Fig. 1*A*). We demonstrated that mutations of critical residues that build the dimer interface impair the quaternary arrangement. The Qua1 domain is highly conserved within the STAR family, with the exception of the loop region, implying that the Qua1 dimeric structures of these proteins are similar.

In fact, while this manuscript was in preparation, the crystal structure of the Gld-1 Qua1 dimer (PDB 3K6T) was reported (49) with the same topology described here for Sam68. The two structures superimpose very well (backbone coordinate r.m.s.d. 1.1 Å; supplemental Fig. 7), although local differences are seen for intermolecular contacts in the loop region, presumably linked to its different composition.

Our results demonstrate that the Qua1 domain alone is sufficient for dimerization of the Sam68 STAR domain. Although we cannot strictly exclude additional contributions from the KH-Qua2 domain to the dimer interface (25), it seems likely that the Qua1 domain is the main determinant for the dimerization of Sam68. In any case, our results unequivocally show that the Qua1 domain is required for the functional activity in splicing regulation of Sam68 *in vivo*. Thus, dimerization of Sam68 is at least as important for splicing as the binding to the target RNA itself, which is mediated by the KH-Qua2 domain. As the surface of Qua1 is negatively charged (Fig. 3*D*), it is unlikely that the Qua1 domain contributes directly to RNA binding. Instead, Qua1 homodimerization may bring together

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

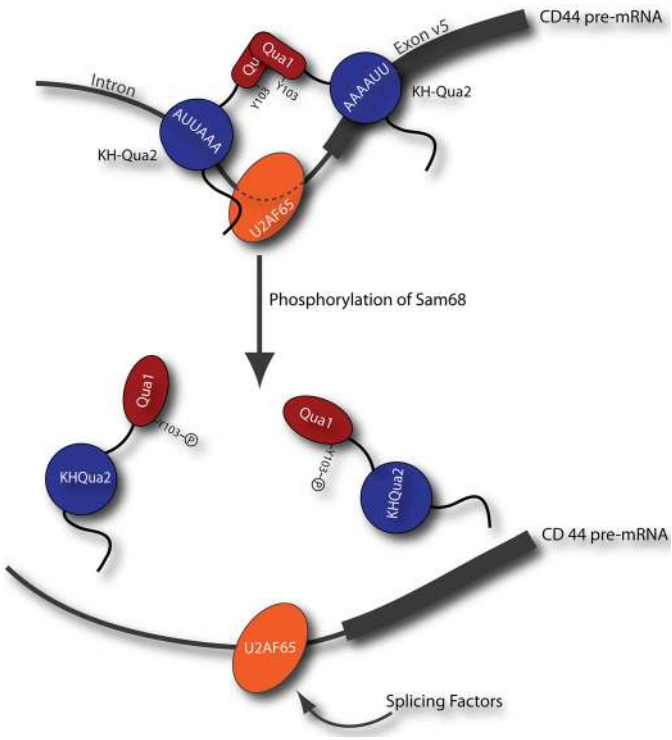


FIGURE 6. **Model of the Sam68 interaction with CD44 pre-mRNA and U2AF65.** Sam68 dimerization via the Qua1 domain may allow simultaneous binding to cognate RNA elements in the CD44 exon v5 and the preceding intron. Phosphorylation of Sam68 has been shown to reduce the RNA binding affinity and thus releases Sam68 from the pre-mRNA. This is required for recruiting constitutive splicing factors and spliceosome assembly. Phosphorylation of Tyr-103 might contribute to the release of Sam68 from the pre-mRNA by destabilizing the dimerization via Qua1.

two KH-Qua2 domains for recognition of bipartite RNA sequences (Fig. 6), a conserved feature within the STAR family.

With respect to alternative splicing of the CD44 v5 exon, Qua1 dimerization may serve two functions. Firstly, it could stabilize binding to two independent RNA binding sites, one within exon v5 and one in the preceding intron, thus promoting splice site definition by other factors. For instance, Sam68 stabilizes binding of U2AF65 to the CD44 pre-mRNA (Fig. 6), which has no canonical U2AF65 binding site downstream of the branch point RNA sequence (9). Secondly, Sam68 appears to promote rearrangement of the spliceosome by leaving the protein-RNA complex. It has been postulated that this is facilitated by a reduced RNA binding affinity linked to phosphorylation (9).

Phosphorylation of Tyr-103 might act as a possible switch in disassembling the Sam68 dimer and thus releasing the protein from the CD44 pre-mRNA (Fig. 6). The Y103S mutation in Sam68 only partially affects the structural integrity *in vitro* (Fig. 4*B*, supplemental Fig. 5), whereas our *in vivo* splicing assays indicate that Tyr-103 is critical for Sam68 function as it strongly reduces exon v5 inclusion on the CD44 mRNA. The loss of splicing activity is comparable with the F118S mutation, which completely destabilizes the dimeric fold. This cannot be explained exclusively by the modest effect of the Y103S mutation on dimerization of Qua1 *in vitro*. One possible explanation for this paradox might be the potential phosphorylation of Tyr-103 by a tyrosine kinase (50). Because Tyr-103 is located close to the edge of the dimer interface, it is in principle accessible to phosphorylation by a kinase. Phosphorylation could sterically interfere with the dimerization. Additionally, the intermolecular hydrogen-bond network, which we have shown to be crucial for dimerization, could be disturbed by additional hydrogen-bond acceptors and donors provided by the phosphate group. Finally, the additional negative charge of the phosphate might lead to repulsion of the two monomers and thereby destabilize the dimer. Phosphorylation of Tyr-103 might thus play a role for the Sam68 splicing activity, beyond this residue being important for the stability of the Qua1 dimer. Future studies should focus on understanding the molecular mechanisms of the function of the Sam68 STAR domain and its modulation by phosphorylation.

Acknowledgments—We thank Harald König, Karlsruhe Institute of Technology (KIT), for providing Sam68 expression vectors and CD44 minigene constructs and Serge Storz for cloning of the initial Qua1 construct. We thank the Bavarian NMR Centre (BNMRZ) for NMR time.

REFERENCES

1. Taylor, S. J., and Shalloway, D. (1994) *Nature* **368,** 867–871
2. Lukong, K. E., and Richard, S. (2003) *Biochim. Biophys. Acta* **1653,** 73–86
3. Vernet, C., and Artzt, K. (1997) *Trends Genet* **13,** 479–484
4. Najib, S., Martín-Romero, C., González-Yanes, C., and Sánchez-Margalet, V. (2005) *Cell Mol. Life Sci.* **62,** 36–43
5. Pillay, I., Nakano, H., and Sharma, S. V. (1996) *Cell Growth Differ.* **7,** 1487–1499
6. Lin, Q., Taylor, S. J., and Shalloway, D. (1997) *J. Biol. Chem.* **272,** 27274–27280
7. Itoh, M., Haga, I., Li, Q. H., and Fujisawa, J. (2002) *Nucleic Acids Res.* **30,** 5452–5464
8. Wang, L. L., Richard, S., and Shaw, A. S. (1995) *J. Biol. Chem.* **270,** 2010–2013
9. Tisserant, A., and König, H. (2008) *PLoS One* **3,** e1418
10. Guitard, E., Barlat, I., Maurier, F., Schweighoffer, F., and Tocque, B. (1998) *Biochem. Biophys. Res. Commun.* **245,** 562–566
11. Babic, I., Jakymiw, A., and Fujita, D. J. (2004) *Oncogene* **23,** 3781–3789
12. Busà, R., Paronetto, M. P., Farini, D., Pierantozzi, E., Botti, F., Angelini, D. F., Attisani, F., Vespasiani, G., and Sette, C. (2007) *Oncogene* **26,** 4372–4382
13. Matter, N., Herrlich, P., and König, H. (2002) *Nature* **420,** 691–695
14. Paronetto, M. P., Achsel, T., Massiello, A., Chalfant, C. E., and Sette, C. (2007) *J. Cell Biol.* **176,** 929–939
15. Arch, R., Wirth, K., Hofmann, M., Ponta, H., Matzku, S., Herrlich, P., and Zöller, M. (1992) *Science* **257,** 682–685
16. Cooper, D. L., and Dougherty, G. J. (1995) *Nat. Med.* **1,** 635–637
17. Sherman, L., Sleeman, J., Dall, P., Hekele, A., Moll, J., Ponta, H., and Herrlich, P. (1996) *Curr. Top. Microbiol. Immunol.* **213,** 249–269
18. Sherman, L., Wainwright, D., Ponta, H., and Herrlich, P. (1998) *Genes Dev.* **12,** 1058–1071
19. Weg-Remers, S., Ponta, H., Herrlich, P., and König, H. (2001) *EMBO J.* **20,** 4194–4203
20. Berglund, J. A., Chua, K., Abovich, N., Reed, R., and Rosbash, M. (1997) *Cell* **89,** 781–787
21. Burd, C. G., and Dreyfuss, G. (1994) *Science* **265,** 615–621
22. Liu, Z., Luyten, I., Bottomley, M. J., Messias, A. C., Houngninou-Molango, S., Sprangers, R., Zanier, K., Krämer, A., and Sattler, M. (2001) *Science* **294,** 1098–1102
23. Maguire, M. L., Guler-Gane, G., Nietlispach, D., Raine, A. R., Zorn, A. M., Standart, N., and Broadhurst, R. W. (2005) *J. Mol. Biol.* **348,** 265–279
24. Zorn, A. M., and Krieg, P. A. (1997) *Genes Dev.* **11,** 2176–2190
25. Chen, T., Damaj, B. B., Herrera, C., Lasko, P., and Richard, S. (1997) *Mol.*

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

Cell. Biol. **17,** 5707–5718
26. Di Fruscio, M., Chen, T., and Richard, S. (1999) *Proc. Natl. Acad. Sci. U.S.A.* **96,** 2710–2715
27. Wu, J., Zhou, L., Tonissen, K., Tee, R., and Artzt, K. (1999) *J. Biol. Chem.* **274,** 29202–29210
28. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) *J. Biomol. NMR* **6,** 277–293
29. Goddard, T., and Kneller, D. (2007) *SPARKY 3*, University of California, San Francisco
30. Jung, Y. S., and Zweckstetter, M. (2004) *J. Biomol. NMR* **30,** 11–23
31. Sattler, M., Schleucher, J., and Griesinger, C. (1999) *Prog. Nucl. Magn. Reson. Spectrosc.* **34,** 93–158
32. Otting, G., and Wüthrich, K. (1990) *Q. Rev. Biophys.* **23,** 39–96
33. Farrow, N. A., Muhandiram, R., Singer, A. U., Pascal, S. M., Kay, C. M., Gish, G., Shoelson, S. E., Pawson, T., Forman-Kay, J. D., and Kay, L. E. (1994) *Biochemistry* **33,** 5984–6003
34. Yang, D., Venters, R. A., Mueller, G. A., Choy, W. Y., and Kay, L. E. (1999) *J. Biomol. NMR* **14,** 333–343
35. Rückert, M., and Otting, G. (2000) *J. Am. Chem. Soc.* **122,** 7793–7797
36. Madl, T., Bermel, W., and Zangger, K. (2009) *Angew. Chem. Int. Ed Engl.* **48,** 8259–8262
37. Güntert, P. (2009) *Eur. Biophys. J.* **38,** 129–143
38. Nilges, M. (1993) *Proteins* **17,** 297–309
39. Shen, Y., Delaglio, F., Cornilescu, G., and Bax, A. (2009) *J. Biomol. NMR* **44,** 213–223
40. Linge, J. P., Williams, M. A., Spronk, C. A., Bonvin, A. M., and Nilges, M. (2003) *Proteins* **50,** 496–506
41. Rieping, W., Habeck, M., Bardiaux, B., Bernard, A., Malliavin, T. E., and Nilges, M. (2007) *Bioinformatics* **23,** 381–382
42. Lipsitz, R. S., and Tjandra, N. (2004) *Annu. Rev. Biophys. Biomol. Struct.* **33,** 387–413
43. Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996) *J. Biomol. NMR* **8,** 477–486
44. Hooft, R. W., Vriend, G., Sander, C., and Abola, E. E. (1996) *Nature* **381,** 272
45. DeLano, W. L. (2002) *The PyMOL Molecular Graphics System*, DeLano Scientific LLC, San Carlos, CA
46. Hadian, K., Vincendeau, M., Mäusbacher, N., Nagel, D., Hauck, S. M., Ueffing, M., Loyter, A., Werner, T., Wolff, H., and Brack-Werner, R. (2009) *J. Biol. Chem.* **284,** 33384–33391
47. Krissinel, E., and Henrick, K. (2004) *Acta Crystallogr. D. Biol. Crystallogr.* **60,** 2256–2268
48. Li, J., Liu, Y., Park, I. W., and He, J. J. (2002) *J. Virol.* **76,** 4526–4535
49. Beuck, C., Szymczyna, B. R., Kerkow, D. E., Carmel, A. B., Columbus, L., Stanfield, R. L., and Williamson, J. R. (2010) *Structure* **18,** 377–389
50. Richard, S., Yu, D., Blumer, K. J., Hausladen, D., Olszowy, M. W., Connelly, P. A., and Shaw, A. S. (1995) *Mol. Cell. Biol.* **15,** 186–197
51. Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2002) in *Current Protocols in Bioinformatics* (Baxevanis, A. D., ed) John Wiley & Sons, Inc., New York
52. Gouet, P., Courcelle, E., Stuart, D. I., and Métoz, F. (1999) *Bioinformatics* **15,** 305–308
53. Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. (2001) *Proc. Natl. Acad. Sci. U.S.A.* **98,** 10037–10041

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

OPEN ACCESS Freely available online

PLOS ONE

Structural and Functional Analysis of the DEAF-1 and BS69 MYND Domains

Fatiha Kateb[1,2], Helene Perrin[3], Konstantinos Tripsianes[1,2], Peijian Zou[1,2,4], Roberta Spadaccini[5], Matthew Bottomley[6], Titus M. Franzmann[2], Johannes Buchner[2], Stephane Ansieau[3], Michael Sattler[1,2,4]*

1 Institute of Structural Biology, Helmholtz Zentrum München, Neuherberg, Germany, **2** Biomolecular NMR and Center for Integrated Protein Science Munich at Department Chemie, Technische Universität München, Garching, Germany, **3** Institut National de la Santé Et de la Recherche Médicale U590, Centre Léon Bérard, Université Claude Bernard Lyon I, Lyon, France, **4** Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, China, **5** Dipartimento di Chimica, Università degli Studi di Napoli "Federico II", Napoli, Italy, **6** Novartis Vaccines and Diagnostics, Siena, Italy

Abstract

DEAF-1 is an important transcriptional regulator that is required for embryonic development and is linked to clinical depression and suicidal behavior in humans. It comprises various structural domains, including a SAND domain that mediates DNA binding and a MYND domain, a cysteine-rich module organized in a Cys_4-Cys_2-His-Cys (C4-C2HC) tandem zinc binding motif. DEAF-1 transcription regulation activity is mediated through interactions with cofactors such as NCoR and SMRT. Despite the important biological role of the DEAF-1 protein, little is known regarding the structure and binding properties of its MYND domain. Here, we report the solution structure, dynamics and ligand binding of the human DEAF-1 MYND domain encompassing residues 501–544 determined by NMR spectroscopy. The structure adopts a $\beta\beta\alpha$ fold that exhibits tandem zinc-binding sites with a cross-brace topology, similar to the MYND domains in AML1/ETO and other proteins. We show that the DEAF-1 MYND domain binds to peptides derived from SMRT and NCoR corepressors. The binding surface mapped by NMR titrations is similar to the one previously reported for AML1/ETO. The ligand binding and molecular functions of the related BS69 MYND domain were studied based on a homology model and mutational analysis. Interestingly, the interaction between BS69 and its binding partners (viral and cellular proteins) seems to require distinct charged residues flanking the predicted MYND domain fold, suggesting a different binding mode. Our findings demonstrate that the MYND domain is a conserved zinc binding fold that plays important roles in transcriptional regulation by mediating distinct molecular interactions with viral and cellular proteins.

Citation: Kateb F, Perrin H, Tripsianes K, Zou P, Spadaccini R, et al. (2013) Structural and Functional Analysis of the DEAF-1 and BS69 MYND Domains. PLoS ONE 8(1): e54715. doi:10.1371/journal.pone.0054715

Editor: Ramón Campos-Olivas, Spanish National Cancer Center, Spain

Received August 12, 2012; **Accepted** December 4, 2012; **Published** January 25, 2013

Copyright: © 2013 Kateb et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: F.K and K.T. acknowledge support by the Alexander von Humboldt foundation. R.S. and M.J.B. are grateful for long-term EMBO and Marie-Curie fellowships, respectively. P.Z. is supported by the grant (KSCX2-YW-BR-4) from the Chinese Academy of Science. M.S. is supported by the Deutsche Forschungsgemeinschaft. M.S. holds a Visiting Professorship for Senior International Scientists at the Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: M.J. Bottomley is currently an employee of Novartis Vaccines & Diagnostics. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: sattler@helmholtz-muenchen.de

Introduction

Initiation of transcription is tightly regulated by a complex interplay of transcription factors which can either directly bind to the DNA or to other transcriptional co-factors. Mutations in these proteins are linked to cancer and other pathologies. DEAF-1 is an important transcriptional regulator that is required for embryonic development [1–3] and linked to clinical depression and suicidal behavior in humans [4]. Michelson et al, showed that deletion of the MYND domain in human DEAF-1 results in a protein less effective than the full length protein in transcriptional repression of nuclear ribonucleoprotein A2/B1 Promoter [2]. DEAF-1 has also been identified as a protein partner of LMO4 that plays important roles in mammary gland development and breast cancers. Consistently, overexpression of DEAF-1 *in vitro* as well as *in vivo* was found to promote mammary epithelial proliferation [5]. Furthermore, DEAF-1 deficient mice displayed phenotypic abnormalities similar to those observed in LMO4 mutants including exencephaly, transformation of cervical segments, and rib cage abnormalities [6]. Through LMO4, DEAF-1 also interacts with the tumor suppressor BRCA1, potentially linking DEAF-1 to breast cancer development [2,7,8].

The DEAF-1 protein contains two conserved domains, the SAND (Sp100, AIRE-1, NucP41/75 and DEAF-1) domain, originally named KDWK after the conserved amino acid core, and a cysteine rich region called the MYND (MYeloid translocation protein 8, Nervy and DEAF-1) domain [1]. To the best of our knowledge, it is the only known mammalian protein that contains both domains. Transcriptional regulation by DEAF-1 involves recognition of target DNA sequences comprising TTCG [1,3] or TTCGG [9] motifs in Drosophila and humans, respectively by the SAND domain [1], which was previously found to adopt a novel DNA binding fold [10]. The MYND domain of DEAF-1 failed to specifically bind DNA, but may still play a role in transcriptional regulation by interacting with corepressor proteins. For example, in the chimeric AML1/ETO protein produced by the 8;21 translocation associated with 12–15% of acute myeloid leukemias, the ETO MYND domain binds

the SMRT and NCoR corepressors [11–13], and MYND domain of the adenoviral protein BS69 was shown to interact with NCoR [14]. BS69 also interacts with the adenoviral oncoprotein E1A as well as with the EBNA2 (Epstein-Barr virus-induced Nuclear Antigen2) protein. These interactions depend on the presence of a PXLXP amino acid motif (X: any amino acid). Interestingly, whereas MYND domains in the Bra-1 and Bra-2 proteins were found to bind E1A and EBNA2, no interaction of the RACK7 or ETO MYND domains with these viral proteins was observed [15], suggesting different binding specificities among MYND domains.

MYND domains comprise 40 to 60 residues with two conserved zinc-binding motifs that are often located at the amino or carboxyl termini of the corresponding full length proteins. MYND domains have two consensus zinc-binding motifs C-X-X-C and C/H-X-X-X-C with a characteristic conserved spacing of cysteine and histidine residues. Structures of the MYND domains of AML1/ETO [13] and in several paralogs of SMYD proteins have been reported recently [16–23] and demonstrated that the tandem zinc-binding motifs are organized in cross-brace topology similar to that observed in RING (Really Interesting New Gene-1) finger domains [24]. The structure of ETO MYND bound to a peptide motif in the SMRT co-repressor revealed molecular details for the recognition of a proline-rich PPPLI motif in SMRT and NCoR corepressors [13]. An intact zinc coordination by the SMYD3 MYND domain was found to be required for an interaction with the NCoR corepressor in coimmunoprecipitation experiments [22], while another report suggested that the SMYD3 MYND domain could be involved in DNA binding [17]. However, based on the conservation of MYND domain residues that are important for corepressor binding by ETO it seems plausible that SMYD MYND domains may also bind SMRT and NCoR corepressors. However, this has not been demonstrated experimentally.

Despite the important biological role of the DEAF-1 protein little is known regarding the structure and binding properties of its MYND domain. Here, we report the solution structure, backbone dynamics and ligand binding properties of the human DEAF-1 MYND domain encompassing residues 501–544 using NMR spectroscopy. The structure presented herein corrects our previous report [25,26] of an incorrect structure of the DEAF-1 MYND. Compared to the previous study where a longer construct of the DEAF-1 MYND domain was used the present study is based on a shorter construct that yields significantly improved NMR data. The structure of the DEAF-1 MYND domain reveals a tandem zinc-binding fold with a cross-brace topology, similar to other MYND domains. We show that DEAF-1 MYND binds to peptides derived from SMRT and NCoR and have mapped the binding surface by using NMR titrations. In addition, we have characterized the ligand binding properties and molecular functions of the related BS69 MYND domain based on a homology model and mutational analysis. Our findings demonstrate that the conserved MYND domain fold found in different transcriptional regulators is a versatile scaffold that may support distinct molecular interactions with viral and cellular proteins.

Results

Zinc binding of the DEAF-1 MYND domain

For structural and biochemical studies various constructs of the DEAF-1 MYND domain (Figure 1a) were cloned and expressed in *E. coli*. A protein comprising residues 501–544 of human DEAF-1 representing the globular fold of the MYND domain was used for further studies. The tandem zinc fingers of MYND domains bind two zinc ions in a tetrahedral coordination via seven cysteine and one histidine side chain. The dependence of the MYND domain fold on zinc coordination was confirmed by the addition of 5 mM EDTA to the DEAF-1 MYND domain. The addition of this metal cation chelator results in a poorly dispersed NMR spectrum, particularly in the ^{1}H chemical shift region around 6–9 ppm, characteristic of an unfolded protein (Figure S1a). This confirms the importance of zinc ions in maintaining the fold, as has been reported for other MYND domains [13].The DEAF-1 MYND domain contains seven cysteines that are all involved in zinc chelation and are identified based on the sequence conservation (Figure 1a). We used NMR to identify which of the three histidines is involved in zinc coordination and characterized their tautomeric states. Different patterns were observed in a long-range $^{1}H,^{15}N$ HSQC experiment correlating histidine $H^{\varepsilon 1}$ and $H^{\delta 2}$ protons to their neighboring nitrogen $N^{\delta 1}$ and $N^{\varepsilon 2}$ atoms [27]. Figure 1b shows that $N^{\delta 1}$ and $N^{\varepsilon 2}$ of His519 have very similar chemical shifts, indicating a doubly protonated state, where the charge is distributed between the two nitrogen atoms. In contrast the imidazole rings of histidines H536 and H538 are protonated at their $N^{\delta 1}$ and $N^{\varepsilon 2}$ nitrogens, respectively. Binding to a zinc ion usually involves $N^{\varepsilon 2}$, but coordination via $N^{\delta 1}$ has also been observed in some RING fingers [24,28]. On the basis of the tautomeric states we can thus not exclude that H538 might mediate the zinc coordination. However, mutation of His538 to Ser results in a well dispersed spectrum that strongly resembles the wild type one (Figure S1b) indicating that H538 is not required for maintaining the fold of the DEAF-1 MYND domain and thus is not involved in zinc coordination. Taken together, these experiments demonstrate that zinc coordination of the DEAF-1 MYND domain involves His536 and the seven cysteine residues.

Solution structure of the DEAF-1 MYND domain

The solution structure of the MYND domain was determined using standard triple resonance NMR experiments [29]. A summary of NMR data and the secondary structure is shown in Figure 1c. Distance and orientational restraints were derived from NOE intensities and residual dipolar couplings, respectively. In addition, backbone torsion angle restraints were defined based on secondary chemical shifts using TALOS+ [30]. Special care was taken to ensure the tetrahedral zinc coordination geometry during the structure calculations by a combination of distance and angle constraints as described in the materials and method section. Structures were calculated using CYANA [31] and further refined using ARIA/CNS [32,33]. Initial calculations in CYANA using only NOE-derived distance restraints and dihedral angle restraints identified the protein fold with the candidate residues for zinc coordination being in close proximity, and consistent with a cross-brace arrangement of the two zinc-binding motifs. Once the topology of the Zn^{2+}-coordinated residues was confirmed, subsequent CYANA structure calculations employed distance restraints that imposed tetrahedral Zn^{2+}-coordination to Cys and His residues. The zinc coordination has been confirmed by the unambiguous assignment of medium and long-range NOEs, for example between C524 H^{N} and C504 $H^{\beta *}$, C524 H^{N} and C528 $H^{\beta *}$ for the first binding site; and between H536 $H^{\varepsilon 1}$ and C540 H^{N}, H536 $H^{\varepsilon 1}$ and C518 $H^{\beta 1}$, H536 $H^{\varepsilon 1}$ and K520 H^{N}, H536 $H^{\varepsilon 1}$ and C515 $H^{\beta 2}$ for the second binding site (Figure 2e). The final CYANA structures were refined in a box of explicit water molecules using ARIA/CNS [34], adapted to ensure tetrahedral zinc coordination geometry during this final refinement. The statistics of structure determination and quality analysis are reported in Table 1. The final ensemble of twenty structures for the DEAF-1 MYND domain is shown in Figure 2. Consistent with ^{13}C secondary chemical shifts (Figure 1c) few regions with secondary structure are observed. The structure of the MYND

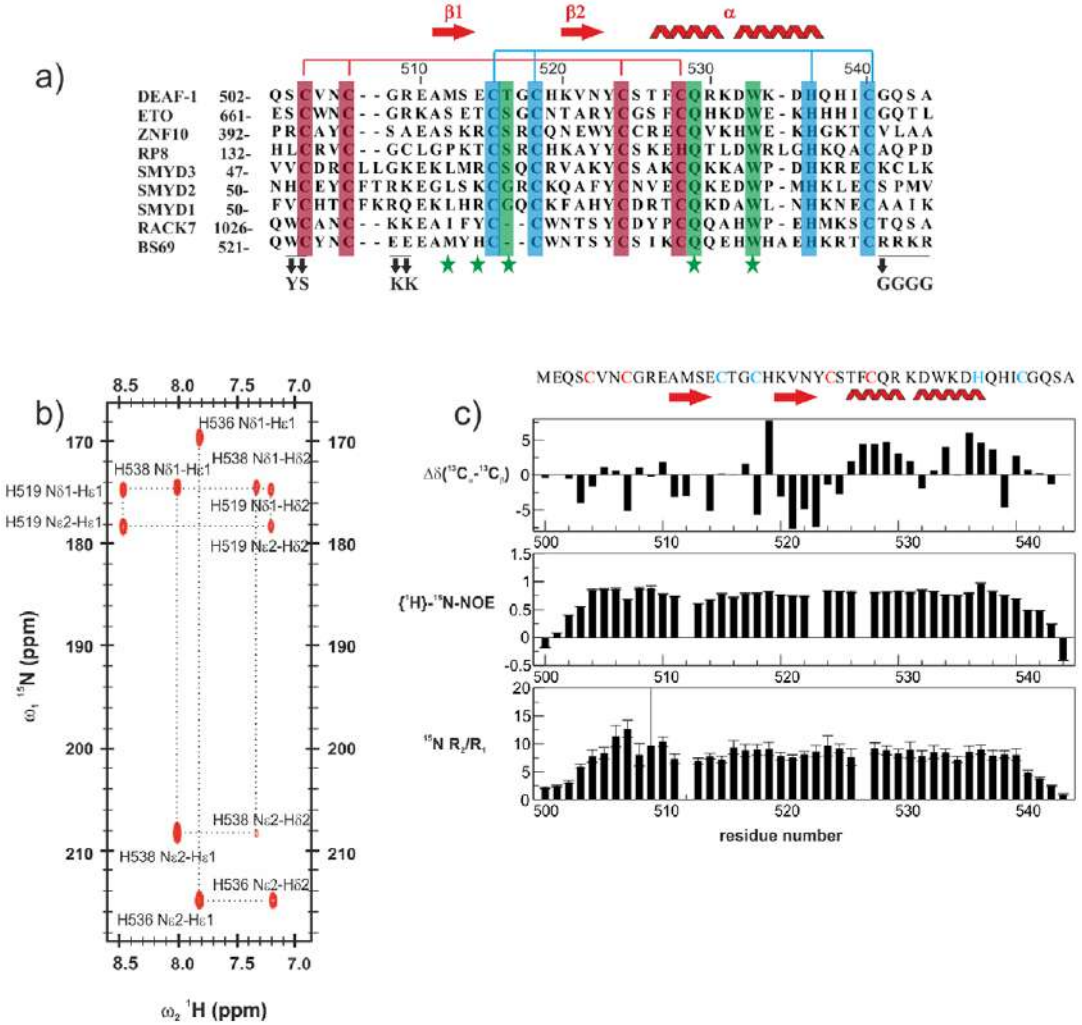


Figure 1. Primary sequence and NMR analysis of the DEAF-1 MYND domain. (a) Sequence alignment of different MYND domains. Residues coordinating the first and second zinc ions are highlighted with red and blue background, respectively. Residues involved in binding to corepressor peptides are indicated with a green star at the bottom, and those interacting through their side chains are highlighted in green. The positions of mutations performed on BS69 are indicated at the bottom. (b) Long range ^{1}H, ^{15}N HSQC spectrum correlating $H^{\varepsilon1}$ and $H^{\delta2}$ to $N^{\delta1}$ and $N^{\varepsilon2}$ through $^{2}J_{HN}$ and $^{3}J_{HN}$ couplings (Pelton et al 1993). The spectrum reveals a different protonation pattern for each histidine sidechain corresponding to the three possible tautomeric states. (c) ^{13}C secondary chemical shifts (top), $\{^{1}H\}$-^{15}N heteronuclear NOE (middle), and ^{15}N R_2/R_1 relaxation rates ratio (bottom) are plotted versus DEAF-1 MYND residue numbers. The secondary structure elements and the amino acid sequence of the protein are indicated at the top of the figure. Residues coordinating the first and second zinc are colored red and blue respectively.
doi:10.1371/journal.pone.0054715.g001

domain is well-defined by characteristic NOEs. It comprises a short β hairpin and an α-helix with a kink around Asp532, which is also consistent with ^{13}C secondary chemical shifts.

The structure of the DEAF-1 MYND domain presents a cross-brace zinc-binding topology (Figure 2), as observed for other MYND domains [13,16,19,20]. The ββα secondary structure and cross-brace zinc-binding topology is shared with RING finger domains, and is common to many extracellular small domains stabilized by disulphide bonds [35]. The first two zinc coordinating cysteine residues in each binding site are located in loop regions, while the last two coordinating residues are within and flanking the α-helix. In the final structures of the DEAF-1 MYND domain the zinc ions show a perfect tetrahedral coordination (Figure 2b). The two zinc atoms are 14 Å far apart, which is another common characteristic of MYND and RING finger domains reflecting the highly conserved spacing between the zinc chelating residues (Figure 1a). The structure of DEAF-1 MYND superimposes very well with other MYND domains, such as those of ETO, ZNF10, SMYD1, SMYD2 and SMYD3 with coordinate root-mean-square deviations for the backbone atoms of 0.99, 1.19, 1.05, 1.07 and 0.99 Å, respectively (Figure 2c). The cross-brace zinc-binding topology of the MYND domain and the arrangement of secondary structure elements strongly resemble RING and PHD finger domains. On the contrary, LIM domains show a sequential zinc-binding topology resulting in two independent zinc-binding sites (Figure 2d).

Backbone dynamics of the DEAF-1 MYND domain

NMR ^{15}N relaxation measurements (^{15}N R_1 and R_2 and $\{^{1}H\}$-^{15}N heteronuclear NOE; Figure 1c) were analyzed using a model-free approach [36,37] as implemented in TENSOR2 [38]. The analysis indicates a tumbling correlation time τ_c of 6.1 ns for the DEAF-1 MYND domain at 22°C. This value is significantly larger than expected for a protein of this size, suggesting possible

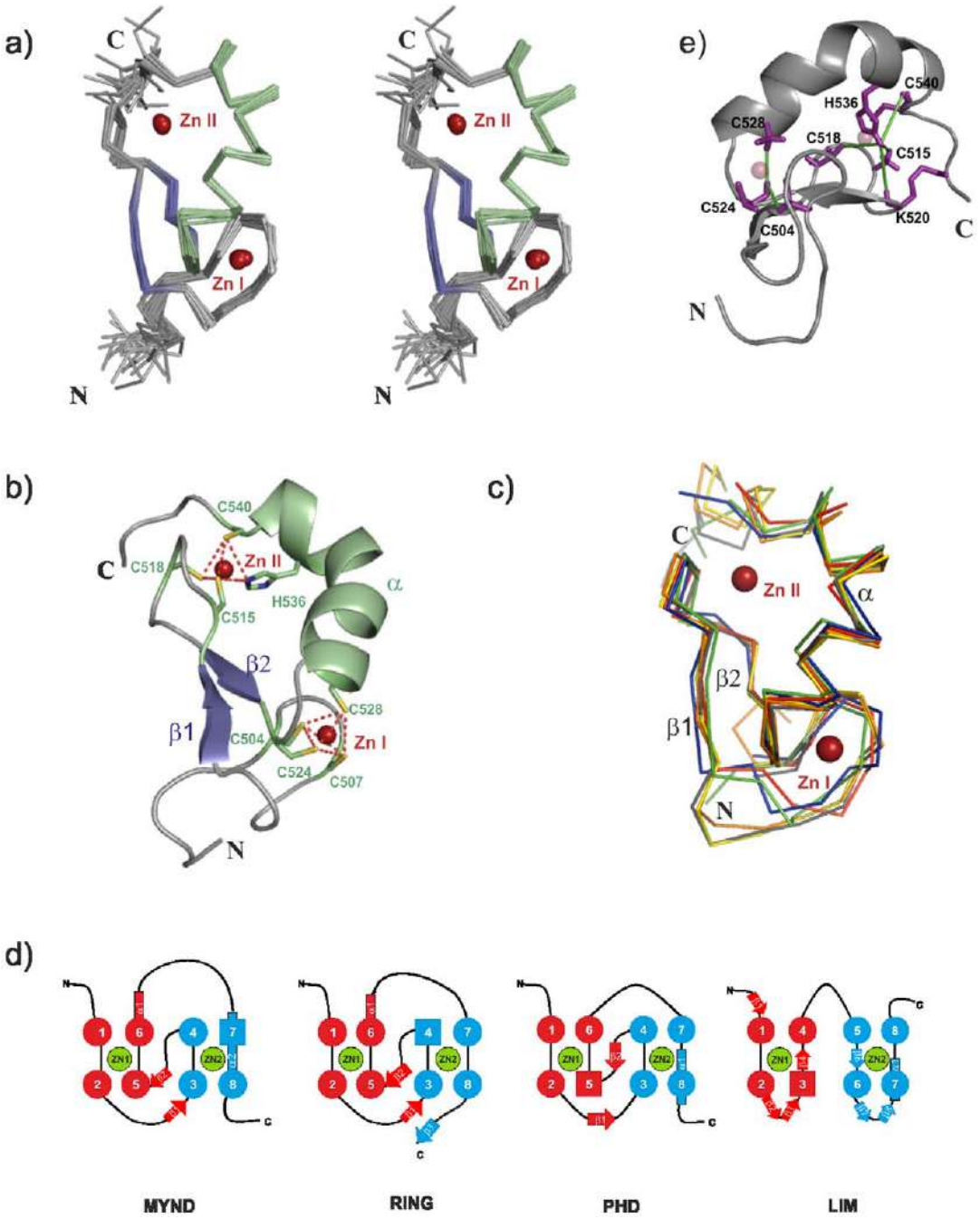


Figure 2. Three-dimensional structure of the DEAF-1 MYND domain. (a) Stereo view of the ensemble of the twenty lowest energy structures of the DEAF-1 MYND domain. α helices and β strands are colored in green and purple respectively, whereas zinc atoms are depicted as red spheres. (b) Ribbon representation of the DEAF-1 MYND domain. Side-chains of residues coordinating the zinc atoms are shown as sticks. The zinc coordination geometry is indicated by red dotted lines. (c) Superposition of DEAF-1 (green), ETO (red), ZNF10 (Blue), SMYD1 (yellow), SMYD2 (orange) and SMYD3 (gray) MYND structures shown in ribbon representation. The two zinc ions are depicted as red spheres. (d) Schematic representation of the zinc-binding pattern and secondary structure elements in MYND, RING, PHD and LIM domains. (e) Cartoon representation of DEAF1-MYND domain. Side chains of residues for which medium and long-range NOEs are observed that unambiguously define the cross-brace zinc binding topology are shown in magenta. Green lines indicate NOEs between C524 H^N/C504 $H^{\beta*}$, C524 H^N/C528 $H^{\beta*}$ for the first binding site; and H536 $H^{\varepsilon 1}$/C540 H^N, H536 $H^{\varepsilon 1}$/C518 $H^{\beta 1}$, H536 $H^{\varepsilon 1}$/K520 H^N, and H536 $H^{\varepsilon 1}$/C515 $H^{\beta 2}$ for the second binding site.
doi:10.1371/journal.pone.0054715.g002

oligomerization of the MYND domain in solution, as has been reported for other MYND domains [13]. Oligomerization is however thought not to be biologically relevant but rather is associated with the high concentration used for the NMR experiments. Analytical ultracentrifugation (AUC) data indicate that the dimerization dissociation constant is 0.5 ± 0.1 μM (Fig. S2a). We identified the residues involved in the oligomerization based on NMR relaxation data and chemical shift changes observed in NMR spectra recorded at different sample concentrations (Fig. S2c). Amide NMR signals of residues located in the

Table 1. Structural statistics of the DEAF-1 MYND domain.

NOE-based distance restraints[1]	
short-range, |i-j|< = 1	439
medium-range, 1<|i-j|<5	180
long-range, |i-j|> = 5	183
Total	802
Dihedral angle restraints[2]	
Φ+Ψ angles	62
Residual dipolar coupling restraints[3]	
H^NN+NC′+H^NC′+H^α C^α	63
Coordinate RMSD for residues 502–541(Å)	
Backbone	0.36±0.10
Heavy atoms	1.10±0.18
Consistency	
RDC Q-factor	0.09±0.002
Ramachandran plot statistics (%)[4]	
Most favoured regions	80.3
Allowed regions	18.9
Generously allowed regions	0.8
Disallowed regions	0.0

Statistics are given for the 20 lowest energy structures after water refinement out of 100 calculated.
[1]Distance restraints were derived from NOE peak intensities using CYANA [31], and then introduced as unambiguous distances in CNS. No distance restraint was violated by more than 0.5 Å.
[2]Torsion angles were predicted using TALOS+ [30]. No dihedral angle restraint was violated by more than 5°.
[3]RDC restraints were incorporated using a harmonic potential. Force constants of 0.2, 0.1, 0.3, and 0.6 kcal mol^{-1} Hz^{-2} for H^HN, NC′, H^NC′ and H^α C^α respectively, were used to reflect the estimated error in the measurement.
[4]Ramachandran plot statistics were obtained using PROCHECK [64] for residues 502–541.
doi:10.1371/journal.pone.0054715.t001

vicinity of the first zinc-binding site (Asn506, Cys507, Arg509, Glu510) exhibit slightly increased R_2/R_1 ratios (Figure 1c, middle) compared to the values for other residues in the structured region of the MYND domain. This suggests the presence of motions in the micro- to milliseconds time scale range, characteristic of exchange phenomena between different conformations, such as monomeric and dimeric states of the protein. Comparison of ^{1}H,^{15}N HSQC spectra acquired at different concentrations show line-broadening of certain peaks even at concentrations as low as 10 μM with some resonances broadened beyond detection (Fig. S2b), consistent with the dimerization constant obtained from analytical ultracentrifugation data (Fig. S2a). ^{1}H,^{15}N HSQC experiments recorded at different salt concentrations (Fig. S3a) and pH (Fig. S3b), as well as ^{15}N relaxation rates (Fig. S3c) show that these buffer conditions do not affect the oligomerization. Most of the residues that exhibit the concentration-dependent line-broadening are located in the first zinc-binding site and thereby identify the dimerization surface. Given that this dimerization interface is far away from the binding site of the MYND domain with co-repressors (see below; Fig. S2c) it can be safely excluded that the dimerization observed *in vitro* could interfere with ligand binding by the MYND domain (see Supplementary Information; Figs. S2, S3 for further details).

Protein-protein interactions mediated by the DEAF-1 MYND

Although the DEAF-1 MYND domain adopts a ββα fold common to many DNA binding zinc finger proteins, it failed to interact specifically with DNA [1]. Rather DNA binding of DEAF-1 occurs through its SAND domain [1]. In contrast, MYND domains are thought to mediate protein-protein interaction [1,39] linked to the recruitment of corepressors [3]. In fact, this has been reported for the ETO MYND domain, which binds the SMRT and NCoR corepressors [12,13,40] through a common PPPLI motif [13]. The high degree of sequence and structural homology between ETO and DEAF-1 MYND domains suggests that the DEAF-1 MYND domain might also interact with SMRT and NCoR. We noted that especially residues important for corepressor binding in ETO are conserved in DEAF-1 (Met512, Glu514, Thr516, Gln529, Trp533; Figure 1a). We therefore used NMR titrations to test whether the DEAF-1 MYND domain does bind peptides derived from SMRT and NCoR. Given the conservation of MYND domain residues we expected that the interaction of the DEAF-1 MYND domain would rely on similar peptide sequences that were reported to bind to ETO [13]. NMR titrations with a 1 mM sample of DEAF-1 MYND and peptides encompassing residues 1111–1120 of SMRT or 1031–1040 of NCoR (Figure 3a) indicate a specific interaction. Binding occurs in the fast-exchange regime on the NMR chemical shift time scale, consistent with a low binding affinity (Figure 3). Notably, DEAF-1 MYND residues that experience large chemical shift changes during the titration (Figure 3a and 3b) are those expected to interact based on the sequence comparison with ETO (see above) or residues located in their vicinity (Figure 3c). The chemical shift perturbations (CSP) observed during titration experiments were fitted to a binding isotherm yielding dissociation constants (K_D) of 5.30±0.54 mM and 3.08±0.12 mM for SMRT and NCoR peptides, respectively. NMR ^{15}N relaxation rates measured in the absence and presence of a large excess of peptide show no significant changes in dynamic properties (data not shown). This confirms that ligand binding does not affect the oligomerization of the domain as expected (see above).

Functional characterization of the BS69 MYND domain

In order to investigate whether the structural characteristics and ligand interactions of MYND domains are conserved we studied the BS69 MYND domain. This domain was previously shown to mediate binding to the NCoR corepressor [14], but also found to interact with cellular partners, such as the Myc-related protein MGA and the viral proteins E1A and EBNA2 depending on the presence of a PXLXP sequence motif in these binding partners [41]. In spite of considerable attempts we were not able to prepare recombinant BS69 MYND in *E. coli* for structural and biochemical studies. However, based on the sequence conservation (~40% identity) between the two domains, we calculated a homology model for the BS69 MYND domain based on the DEAF-1 structure using MODELLER [42–45]. Interestingly, the BS69 MYND model reveals a highly charged surface with a positive face consisting of C-terminal residues on one side, and a negatively charged region on the other side (Figure 4b). The binding of the BS69 MYND domain to MGA, E1A and EBNA2 has been studied by mutational analysis using *in vitro* translated GST-tagged BS69 protein, encompassing the MYND domain. As expected, mutation in one of the zinc coordination sites, completely abolishes the binding to any partners [15], confirming that a structurally intact MYND domain is required for the interaction. Additional mutational analysis was performed using GST pull-down assays. Notably, a charge reversal of Glu527–Glu528

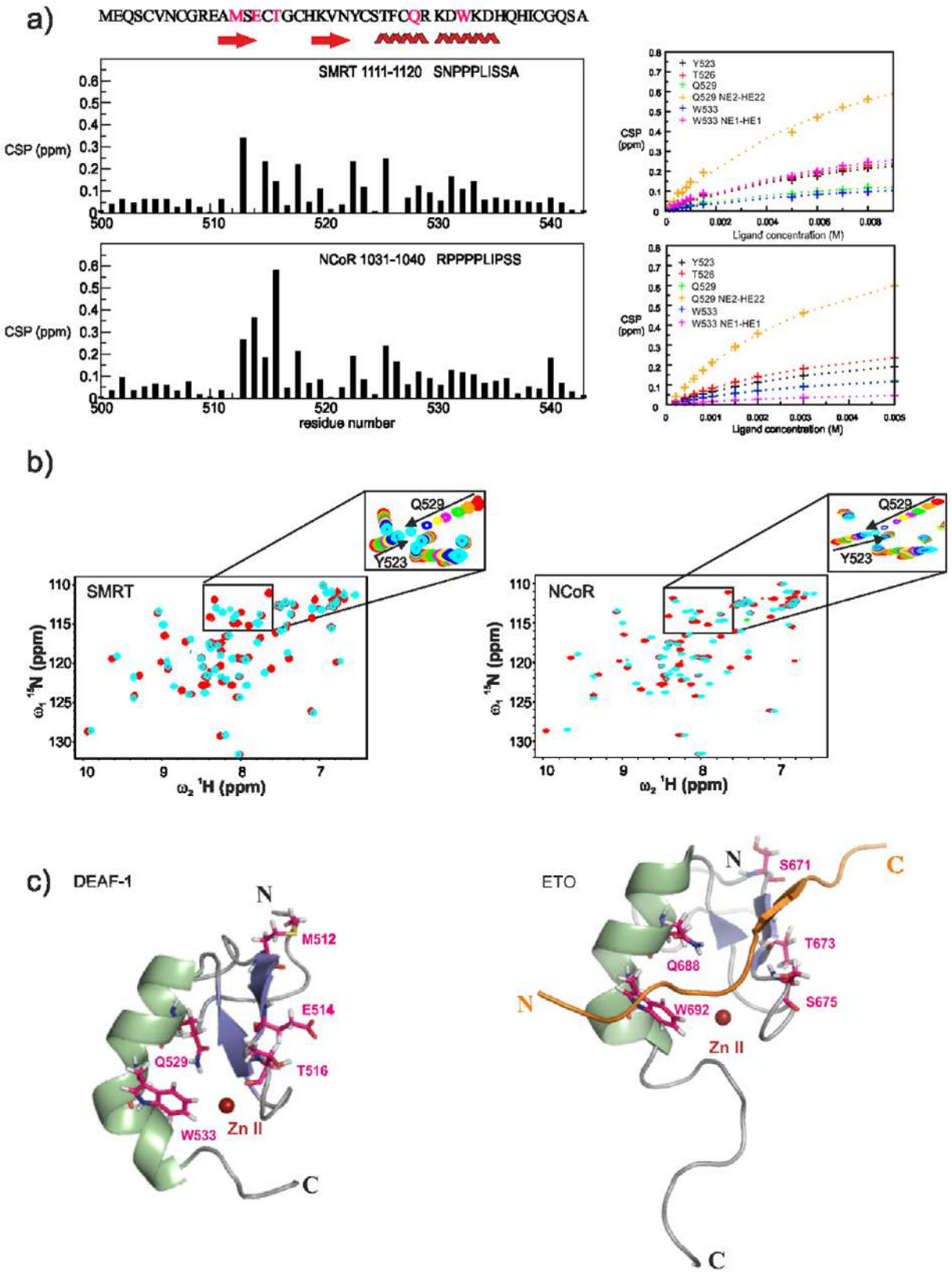


Figure 3. Binding of DEAF-1 MYND to SMRT and NCoR peptides. (a) Chemical Shift Perturbations (CSP, see methods) observed for the interaction between DEAF-1 MYND domain and SMRT (top) and NCoR (bottom) corepressor peptides. The sequence of SMRT and NCoR peptides used for the titrations are indicated in each graph. Secondary structure elements and amino acid sequence of the DEAF-1 MYND domain are shown at the top of the panel. Residues expected to interact with the corepressors are colored magenta. CSPs of residues that are most strongly affected are shown (cross symbols) as a function of ligand concentration for both titration with SMRT (top right) and NCoR (bottom right). The observed CSPs

were fitted to a binding isotherm yielding dissociation constanst of 5.30±0.54 mM and 3.08±0.12 mM for the SMRT and NCoR peptides, respectively. The binding curves are shown as dashed lines. (b) Superposition of ^{1}H, ^{15}N HSQC spectra of a 1 mM sample of the free DEAF-1 MYND domain (red) and upon addition of unlabeled corepressor peptides (cyan) up to a final concentration of 8 mM and 5 mM of SMRT and NCoR peptide (1:8 and 1:5 molar ratio), respectively. The intermediate steps of each titration are zoomed for a sub-region of the corresponding spectrum. In either case binding takes place on the fast exchange regime with respect to the chemical shift time scale. (c) Ribbon representation of DEAF-1 (left) and ETO (right) MYND domains. Residues experiencing the largest chemical shift perturbation upon addition of the corepressor peptides are shown as magenta sticks. In the ETO-SMRT complex structure the corresponding residues are shown as sticks as well, and the SMRT ligand peptide is colored orange and shown in cartoon representation.
doi:10.1371/journal.pone.0054715.g003

(EE527–528KK), involving residues, which are located in a region flanking the first zinc-binding module, strongly reduced the interaction with all three PXLXP ligands (Figure 4c), while the W522Y mutation mainly affects the interaction with the cellular ligand (MGA) but not the viral binding partners (Figure 4d). The BS69 MYND domain contains a set of positively charged residues located at its C-terminus that are absent from other MYND domains, in particular from RACK7, ETO and DEAF-1, which fail to interact with BS69 PXLXP ligands [46]. Mutation of the positively charged residues (RRKR559–562GGGG) abolishes binding to PXLXP ligands (Figure 4d). Individual mutations of these residues further revealed that among these four residues only Arg560 is essential for binding of BS69 to its ligands (Figure 4e; Fig. S5). Taken together the mutational analyses suggest that a positively charged face of the BS69 MYND domain, at the C-terminal end of the zinc-binding fold, is crucial for the interaction with PXLXP ligands and appears to be a unique property of the BS69 protein.

Discussion

The solution structure of the DEAF-1 MYND domain reported here reveals a tandem C4-C2HC zinc-binding motif with a cross-brace topology. This structure corrects our previous report that was based on a C-terminally extended, longer construct of DEAF-1 MYND domain (residues 498–565), for which only few long range NOEs could be obtained (Fig. S4a,b) [25]. For this longer construct key long-range NOEs that unambiguously define the cross-brace zinc topology were only obtained with a 800 MHz NOESY experiment ([26] and Fig. S4a). The structure of the DEAF-1 MYND domain is organized in a ββα fold, which is likely to stabilize the MYND domain fold in the reducing intracellular compartment where disulfide bonds are unstable. The zinc binding topology is somewhat reminiscent of the 1–3, 2–4 disulfide bonds pattern found in extracellular domains, for example in Epidermal Growth Factor (EGF) [47]. Both the zinc coordination and the distance between the two zinc ions are shared by DEAF-1 with other MYND domains as expected from the high sequence similarity observed among this protein family. Note, that an optimized protein construct has been used for structural analysis, which removes any unstructured tails and merely encompasses the structured tandem zinc finger region of the MYND domain. This construct therefore yields high quality NMR spectra and the structural analysis presented herein corrects a previous report [25,26] that had been conducted on a longer construct (see Supplementary Information).

Our NMR titrations using peptides derived from the silencing mediator of retinoid and thyroid receptor (SMRT) and from the nuclear receptor corepressor (NCoR), show that the binding site of DEAF-1 MYND for these corepressors resembles that of the ETO MYND domain. The interaction of these ligands to DEAF-1 MYND is about an order of magnitude weaker than the one previously reported for the ETO MYND domain [13]. This suggests that additional interactions between the full-length DEAF-1 and corepressor proteins may exist that further enhance the interaction. Nevertheless, the striking similarity of the chemical shift perturbations seen for the DEAF-1 and ETO MYND domains and the conservation of residues that are important for the interaction in ETO MYND (Figure 1a) indicates a very similar interaction of the corepressors with the DEAF-1 MYND domain. The recognition of the corepressor peptides by the ETO MYND domain involves hydrogen bonds with backbone or side chains atoms of residues that are also conserved in DEAF-1, including Gln529 [13] (Figure 1a). Another important residue for this interaction is a tryptophan (W533 in DEAF-1; Figure 1a) which packs against a proline in the SMRT ligand peptide bound to the ETO MYND domain [13]. The high degree of similarity between DEAF-1 and ETO MYND domains suggests a similar binding mode, as confirmed by the binding site mapped by our NMR chemical shift perturbations (Figure 3c).

Recently, Foreman and coworkers determined the crystal structure of a multi-domain construct of SMYD3 and found that an intact MYND domain fold is required for interaction with the NCoR corepressor [22]. Although the binding interface with NCoR has not been mapped in this study, the MYND domain residues that are important for corepressor binding are conserved in SMYD3, suggesting a similar binding mode.

Conservation of critical residues involved in the interaction may thus explain why an interaction with corepressors is widely observed among the MYND domain family. While binding to corepressors such as SMRT and NCoR seems to be a common feature of many MYND domains, including those of ETO, BS69, and BOP [12–14,40,48,49], distinct binding specificities exist for other MYND domains, in particular for the interaction with viral proteins such as E1A and EBNA2. The interaction between BS69 and binding partners that comprise a PXLXP sequence motif [41] seems to critically depend on electrostatic contacts. A comparison of the electrostatic surface potentials of the DEAF-1 and BS69 MYND domains (Figure 4a,b, respectively) shows that DEAF-1 MYND is less charged than BS69. This is also indicated by the theoretical pI values of 8.5 and 6.8 for BS69 and DEAF-1 MYND domains respectively, obtained using ExPASy [50]. The less pronounced charge of the DEAF-1 (also seen for the ETO and RACK7 MYND domains, pI's of 6.4 and 6.9 respectively, Figure 1a) seems to correlate with a lack of binding to E1A and EBNA2. In contrast, the interaction of BS69 with these PXLXP-containing proteins depends on the presence of charged residues in the sequence. Since charge reversal of two glutamate residues (EE527–528KK) has a smaller effect, it appears that positively charged arginine and lysine residues flanking the C-terminal end of the zinc-binding fold of the BS69 MYND domain are crucial for the interaction. The positive charges of these side chains could mediate long-range electrostatic interactions with negative charges which may be located in regions flanking the PXLXP motif in the binding partners. The BS69 MYND domain interaction might thus involve a larger binding epitope of which PXLXP merely represents a conserved core motif. Additionally, aromatic residues might contribute to specific PXLXP interactions, by packing against the pyrrolidine ring of the Pro side chains. For example, a tyrosine (Y523) located between the first two zinc-coordinating

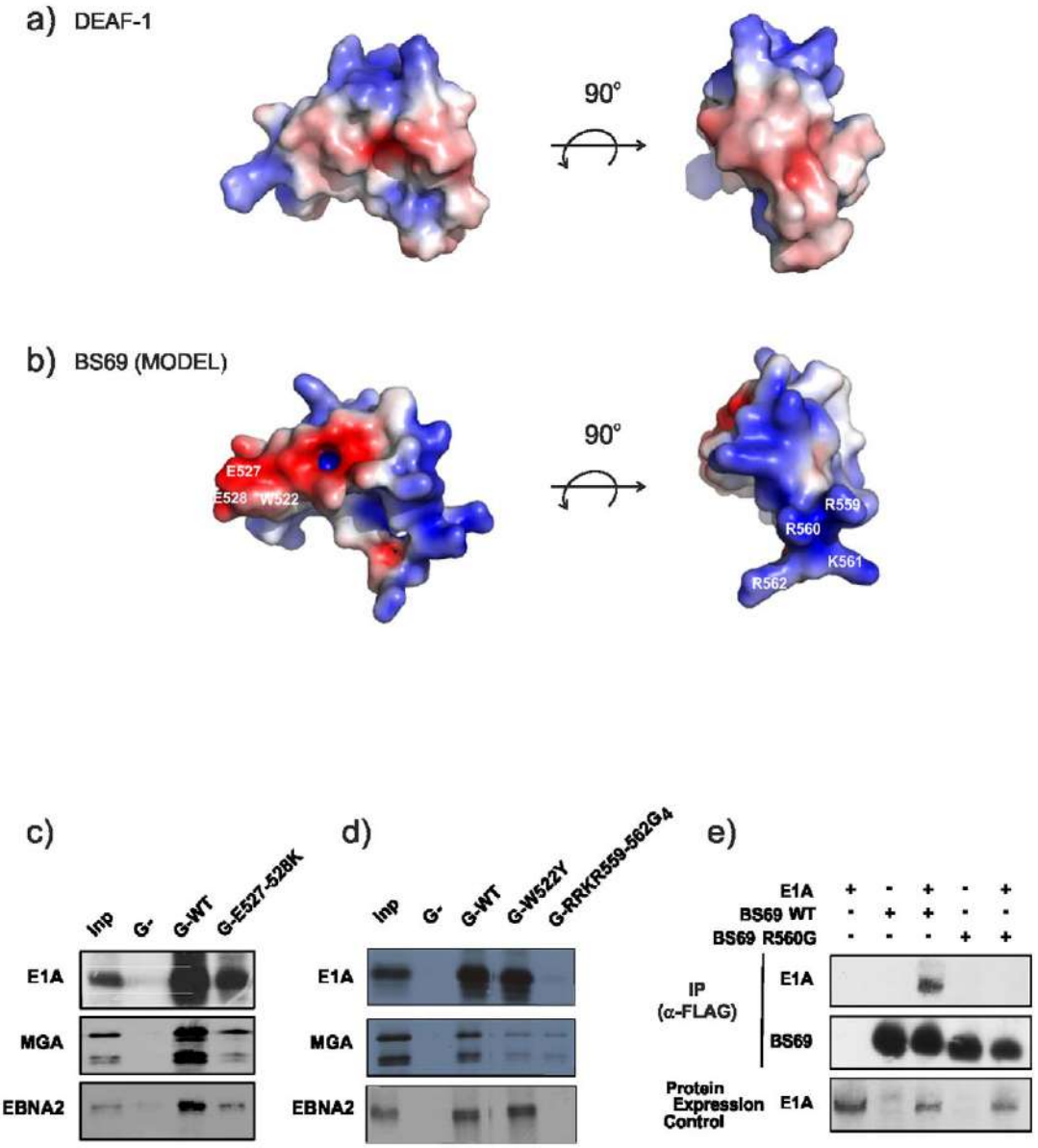


Figure 4. Binding studies and mutational analysis of BS69 MYND domain. (a) Two views of an electrostatic surface representation of the DEAF-1 MYND domain. Positive (blue) and negative (red) electrostatic surface potential is shown at ± 3 k_B T/e^- and was determined using the program APBS [66] in Pymol (www.pymol.org). (b) Corresponding surface views for the BS69 MYND structure obtained from homology modelling. Positive (blue) and negative (red) electrostatic surface potential is displayed at ± 3 k_B T/e^-. The location of the residues that were mutated for binding studies are labelled on the surface representation of the BS69 homology model. (c),(d) Analysis of binding of E1A, EBNA2 and MGA to wild-type or mutant BS69 proteins expressed as a GST-fusion protein or to the GST (G−) moiety as control; Inp: Input 10%; G-RRKR-559-562G4: mutation of residues 559–562 into four glycines. (e) A single point mutation in BS69 abrogates the binding to E1A. QT6 fibroblasts were transfected with 12SE1A and/or FLAG-tagged BS69 expression vectors as indicated. Cellular complexes were immunoprecipitated with an M2 anti-FLAG antibody. Co-immunoprecipitated and ectopic expressions of E1A were revealed by immunoblotting using an M73-E1A antibody.
doi:10.1371/journal.pone.0054715.g004

Cys residues in the BS69 MYND domain only, suggests that hydrophobic interactions of this aromatic residue could further contribute to ligand binding of BS69. This might also explain why ETO MYND domain failed to interact with this PXLXP motif [15].

Although BS69 has been shown to bind to NCoR [14], which also contains a PXLXP recognition motif, the analysis of its sequence reveals that the key residues that are important for corepressor binding (i.e. T516 in DEAF-1, S673 in ETO) are not conserved. This suggests that the binding mode may differ among the MYND domain family. Supporting this hypothesis, the analysis of the BS69 (and RACK7) primary sequence shows that it lacks a Thr/Ser residue and that the two cysteines involved in the second zinc-binding site are directly consecutive without any spacing residues in between (Figure 1a). This likely affects the conformation of this region in the BS69 and RACK7 MYND domains. Additional electrostatic interactions involving residues flanking the MYND fold (Figure 4) could thus additionally contribute to ligand binding by BS69. The NCoR interaction of BS69 may depend critically on electrostatic contacts as observed

for binding to MGA and E1A. In the case of SMYD1, which also lacks the Thr/Ser residue at this position, the binding to a PXLXP motif in the transcriptional activator skNAC depends on additional interactions of amino-terminal residues located in the S-sequence domain [49]. Interestingly, the SMYD1 MYND domain is highly positively charged [19], suggesting that skNAC binding may be enhanced by electrostatics interactions.

Collectively, these data suggest that variations in amino acid sequence drive distinct binding specificities of MYND domains with their cellular and viral partners, even though binding to corepressors such as NCoR and SMRT seems to be a feature that is generally shared by many MYND domains. The relatively weak interaction of DEAF-1 MYND with the corepressor-derived peptides, may be further enhanced by additional protein-protein contacts involving different regions of the full-length DEAF-1 and corepressor proteins. Thus future functional studies of the role of DEAF-1 should consider the possibility that additional domains may further enhance the interaction. The DEAF-1 and ETO MYND domains appear to use the same binding mode when interacting with corepressors, and both fail to bind to MGA, E1A and EBNA2 proteins. In contrast, MYND domains that are more distantly related, such as BS69, may favor electrostatic interactions through their specific charged residues. Future structure/function studies will further highlight the importance of MYND-mediated protein-protein interactions in transcriptional regulation with viral and cellular proteins.

Materials and Methods

Cloning, protein expression and sample preparation

The DNA fragments comprising the human DEAF-1 MYND domain (residues 501–544 and 496–565 for the short and long constructs respectively) plus an N-terminal insertion of three amino acids GAM, were PCR-amplified from Human cDNA. The forward primer: AAACCATGGAGCAGTCCTGCGTTAAC containing *Nco*I site and the reverse primer: AAAGGTAGGT-CATGCTGACTGGCCGCATATG containing *Kpn*I. site were designed to clone MYND domain into the expression vector pETMThx, which was modified from pET24d to include an N-terminal poly-histidine tag, thioredoxin as a fused protein and a TEV cleavage site.

The DEAF-1 MYND constructs were transformed into *E.coli* strain BL21(DE3)(Novagen, USA). To prepare ^{15}N, ^{13}C and/or ^{15}N isotope-labeled proteins expression was carried out in M9 medium containing [U-^{13}C] glucose and/or [U-^{15}N] NH_4Cl as sole carbon and nitrogen sources. Overnight cultures were grown in LB medium at 37°C, diluted 50 fold into the M9 medium, continuously grown at 37°C to an $OD_{600\ nm}$ of 0.6. The culture was induced with 1 mM IPTG. During the induction 0.3 mM $ZnCl_2$ was added and the culture was switched to 20°C for 10–16 hours. All media contained 50 μg/ml kanamycin.

Bacteria cells were harvested by centrifugation at 8,000 g for 30 min and re-suspended in 3 ml/g (wet weight) lysis buffer (50 mM Tris-HCl, pH 8.0, 300 mM NaCl, 5 mM imidazole), supplemented with protease inhibition cocktail complete (Roche, Germany) and DNase I (10 μg/ml). Cells were lysed by pulsed sonification (6 min, 30% power, large probe, Fisher Scientific model 550). Cell debris was removed by centrifugation at 15,000 g for one hour. All proteins were purified by two steps of chromatography. The supernatants were applied to Ni^{2+}-NTA (QIAGEN, Germany), equilibrated with the lysis buffer. The bound proteins were eluted with 400 mM imidazole in the lysis buffer. The N-terminal His-tag and the fusion thioredoxin were removed by adding 1:50 (molar ratio) TEV protease, supplemented with 2 mM DTT to the pooled protein solution. After removing the residual DTT, using a desalting column, PD10 (Amersham Biosciences, Sweden), the TEV protease and cleaved His-tag were removed by an additional Ni-chelating affinity chromatography.

The flow through from the second Ni-chelating affinity chromatography was collected, concentrated and loaded on to a gel filtration HiLoad Superdex 75 16/60 column (Amersham Biosciences, Sweden) previously equilibrated with 50 mM phosphate buffer, pH 6.5, containing 100 mM NaCl, 2 mM DTT. The fractions containing MYND proteins were pooled and concentrated to 0.3–1 mM for NMR measurements.

NMR spectroscopy

The chemical shifts of the DEAF-1 MYND domain were assigned using standard heteronuclear experiments acquired at 295 K on a 1 mM uniformly $^{15}N/^{13}C$ labeled sample in 90% H_2O/10% D_20 [29]. Experiments were carried out on Bruker spectrometers operating at a proton frequency between 500 and 900 MHz. All spectra were processed using the package NMRPipe/NMRDraw [51]. For structure determination, a 2D NOESY was recorded on a 600 MHz spectrometer equipped with a TCI cryo-probehead, whereas ^{15}N- and ^{13}C-edited NOESY spectra were recorded on an Avance 900 spectrometer equipped with a TXI cryo-probehead with a mixing time of 80 ms (100 ms for the aromatic ^{13}C-edited NOESY spectrum). Data were analyzed using Sparky [52].

A 50 mg/ml phage solution was used to weakly align the sample in an anisotropic phase, resulting in a ^2H quadrupolar splitting of ca 20 Hz. H^{α}-C^{α}, H^N-C′, N-C′ and H^N-N residual dipolar couplings (RDCs) were recorded on a 600 MHz spectrometer equipped with a TCI cryoprobehead using experiments described elsewhere [53–55].

NMR relaxation measurements including ^{15}N longitudinal (R_1), transverse (R_2) relaxation rates, and the $\{^1H\}$-^{15}N heteronuclear Overhauser (NOE) effect were carried out on the in-house 750 MHz spectrometer equipped with a TXI probehead using standard methods [56]. For R_1 and R_2 relaxation rates, ten different time-points were recorded in an interleaved manner. Peak intensities were fitted to a two-parameter exponential decay using NMRViewJ [57].

Dissociation constants for the interaction of DEAF-1 MYND with proline-rich peptides derived from SMRT and NCoR corepressors were determined from fitting chemical shift changes monitored in NMR titration experiments to a binding isotherm. A 1 mM solution of ^{15}N-labeled MYND domain was titrated with peptides SNPPPLISSA (SMRT) and RPPPPLIPSS (NCoR) up to a 9-fold and 5-fold molar excess, respectively. The changes in ^1H and ^{15}N chemical shifts are referred to as chemical shift perturbations (CSP) and were calculated as:

$$CSP = \sqrt{\delta_{HN}^2 + \frac{\delta_N^2}{6^2}}$$

The CSPs were fitted to a binding isotherm using the equation:

$$CSP = \frac{CSP_{\max}}{[2P_T]}\left([L] + [P_T] + K_D - \sqrt{([L] + [P_T] + K_D)^2 - 4[P_T][L]}\right)$$

where CSP, is the chemical shift perturbation at a given peptide concentration [L], CSP_{max} is the chemical shift perturbation at

saturation, $[P_T]$ is the total protein concentration, and K_D, the dissociation constant.

Structure calculation

First structure calculations were performed with CYANA [31] using NOE cross-peaks that were automatically assigned and manually checked, combined with distance restraints that impose tetrahedral Zn^{2+}-coordination, torsion angle restraints derived from chemical shifts using TALOS+ [30] and RDC restraints. For the latter, the axial component and rhombicity of the alignment tensor were determined by CYANA 3.0 using a grid search approach, the result being in full agreement with what was obtained from the analysis of the RDC distribution [58]. Distance restraints derived from the CYANA calculations, together with torsion angle and RDC restraints were used in Aria1.2 for a water refinement calculation [59–63]. The zinc coordination geometry was defined and maintained by distance (Zn-S^{γ} 2.3 Å and Zn-$N^{\varepsilon 2}$ 2.0 Å) and angle (S^{γ}-Zn-S^{γ}, $N^{\varepsilon 2}$-Zn-S^{γ}, Zn-S^{γ}-C^{β} 109.5° and Zn-$N^{\varepsilon 2}$-$C^{\delta 2}$ 120°) restraints. This procedure ensures a proper tetrahedral zinc coordination, which is not fulfilled in most of the NMR-derived zinc-binding folds found in the Protein Data Bank. In this final step a total of 100 structures were calculated from which the twenty lowest energy structures were used for quality and structure validation using the iCING web interface (http://nmr.cmbi.ru.nl/icing/) and PROCHECK [64]. The structure of the BS69 MYND domain was generated by comparative modeling with the program MODELLER [42–45] based on sequence alignment using the coordinates of DEAF-1 MYND domain as template. Molecular images were generated using PYMOL (www.pymol.org).

BS69 binding experiments

Mutations were performed using the QuickChange mutagenesis kit (Stratagene). Wild type and mutants BS69 GST- fusion proteins encompass the 150 last C-terminal residues of the protein (BS69Δ431) [65]. GST-protein production, GST pull-down assay, cell transfections and immunoprecipitation experiments have been previously described [41]. For immunoprecipitation experiments, briefly, the BS69 fragment (aa 451–562 R560G) was subcloned as the FLAG-pcDNA3 vector. Combination of BS69 and 12SE1A expression vectors were transiently transfected in quail fibroblasts (QT6). Cells were harvested 24 h post-transfection, lysed and successively incubated with an anti-FLAG M2 antibody (Integra Biosciences) and Protein A sepharose (Amersham Biosciences Inc.). After separation of proteins on SDS-PAGE, BS69 and E1A were visualized with an M2 anti-FLAG antibody and an M73 anti-E1A antibody (Calbiochem) respectively.

Supporting Information

Figure S1 Dependence of the MYND domain fold on zinc coordination. a. Overlay of 1H 1D spectra obtained from the wild type protein in the absence (red) and presence (cyan) of EDTA. b. Overlay of 1H-15N HSQCs of the wild type protein (red) and H538S mutant (cyan).
(TIF)

Figure S2 Oligomerization of DEAF-1 MYND domain. a. Left: Continuous sedimentation distribution obtained for 86 μM DEAF-1 MYND domain. Analysis revealed a 9.95 kDa species suggesting that MYND monomers (5.2 kDa) associate into dimers. The monomer and dimer appear to be in a in rapid monomer-dimer equilibrium (data not shown). Right: Concentration-dependent sedimentation velocity analysis of DEAF-1 MYND domain. Dimerization occurs at sub-micromolar concentrations. The data were fitted to a self-association model yielding an estimated K_D of ~0.5 μM (±0.1 μM). b. Overlay of 1H-15N HSQC spectra recorded at 1 mM (black) and 50 μM (cyan) protein concentration. Cross-peaks that experience line-broadening upon dilution are annotated with the corresponding residue numbers. c. Cartoon representation of the DEAF-1 MYND domain. Residues affected by the dilution of the sample are clustered around the first zinc binding site and are shown as cyan sticks, whereas those involved in binding to co-repressor peptides are shown as magenta sticks. Residue H538 is also shown in black.
(TIF)

Figure S3 Salt and pH dependance of oligomerization of DEAF-1 MYND domain. a. Overlay of 1H-15N HSQC spectra (left) and long-range 1H-15N HSQC spectra (right) recorded at a NaCl concentration of 100 mM (black) and 400 mM (green) b. Overlay of 1H-15N HSQC spectra (left) and long-range 1H-15N HSQC spectra (right) recorded at a pH of 6.5 (black) and 6 (red). At pH 6, H538 tends to a conformation where both nitrogens of the side chain, are protonated, thus excluding any possible ion binding. c. 15N R1 and R2 relaxation rates of DEAF-1 MYND domain for three different samples, i.e., [NaCl] = 100 mM and pH = 6.5 (black), [NaCl] = 400 mM and pH = 6.5 (green), [NaCl] = 100 mM and pH = 6 (red). No significant change in relaxation rates are observed changing the buffer conditions. The oligomerization state revealed by the high R2/R1 ratio is thus likely due to non specific interactions.
(TIF)

Figure S4 Correction of the previously reported structure of DEAF-1 MYND domain. a. Cartoon representation of the corrected DEAF1-MYND structure with long-range NOEs from the new NMR data. Side chains of residues showing long-range NOEs are shown in as magenta sticks. Newly observed long-range NOEs, which define the cross-brace zinc binding topology, are indicated by green lines. Long range NOEs that were also observed previously are shown as dotted black lines. b. Comparison of original and new NOESY experiments showing important NOEs originating from aromatic residues. Left: The NOE contacts of the His 536 Hε1 proton in a homonuclear 2D NOESY (500 MHz, 120 ms mixing time) and in the 3D aromatic 13C edited NOESY spectrum (800 MHz, 300 ms mixing time). Right: The same for NOEs involving the Tyr 523 Hε1 proton. NOEs observed only with the new experiments are green boxed.
(TIF)

Figure S5 Binding of E1A to wild-type or mutant BS69 proteins. The effect of individual mutation of residues RRKR 559–562 of Bs69 for E1A binding was tested. 2× 106 QT6 fibroblasts were transiently transfected with 12SE1A and wt or mutant FLAG-tagged BS69 411–561 peptides as notified on top. 24 h post-transfection, protein lysates were immunoprecipitated with a 10 μg anti-FLAG antibody in a 100 mM NaCl, 20 mM Tris pH8, 0.5% NP40 buffer and Protein A sepharose beads. After extensive washes in the binding buffer, proteins were eluded from beads and separated by SDS-PAGE. Immuno-precipitated E1A protein was revealed by western blotting. Individual mutations of these residues indicate that among these four residues only Arg560 is essential for binding of BS69 to E1A.
(TIF)

Acknowledgments

We thank Alexander Gasch, Bernd Simon and Gunter Stier (EMBL Heidelberg) for help; M. Beyermann (FMP, Berlin) for peptide synthesis and Alexander Bepperling for preliminary analytical ultracentrifugation

experiments. The coordinates and NMR restraint files for the human DEAF-1 MYND domain are deposited in the PDB with accession number 4A24. Chemical shifts are deposited in the BMRB, accession code: 18004.

Author Contributions

Conceived and designed the experiments: FK SA MS. Performed the experiments: FK HP KT PZ RS MB TMF. Analyzed the data: FK HP TMF SA MS. Contributed reagents/materials/analysis tools: JB. Wrote the paper: FK SA MS.

References

1. Gross CT, McGinnis W (1996) DEAF-1, a novel protein that binds an essential region in a Deformed response element. EMBO J 15: 1961–1970.
2. Michelson RJ, Collard MW, Ziemba AJ, Persinger J, Bartholomew B, et al. (1999) Nuclear DEAF-1-related (NUDR) Protein Contains a Novel DNA Binding Domain and Represses Transcription of the Heterogeneous Nuclear Ribonucleoprotein A2/B1 Promoter. JBC 274: 30510–30519.
3. Veraksa A, Kennison J, McGinnis W (2002) DEAF-1 Function Is Essential for the Early Embryonic Development of Drosophila. genesis 33: 67–76.
4. Lemonde S, Turecki G, Bqkish D, Du L, Hrdina PD, et al. (2003) Impaired Repression at a 5-Hydroxytryptamine 1A Receptor Gene Polymorphism Associated with Major Depression and Suicide. J Neurosci 23: 8788–8799.
5. Barker HE, Smyth GK, Wettenhall J, Ward TA, Bath ML, et al. (2008) Deaf-1 regulates epithelial cell proliferation and side-branching in the mammary gland. BMC Dev Biol 8.
6. Hahm K, Sum EY, Fujiwara Y, Lindeman GJ, Visvader JE, et al. (2004) Defective neural tube closure and anteroposterior patterning in mice lacking the LIM protein LMO4 or its interacting partner Deaf-1. Mol Cell Biol 24: 2074–2082.
7. Sum EY, Peng B, Yu X, Chen J, Byrne J, et al. (2002) The LIM domain protein LMO4 interacts with the cofactor CtIP and the tumor suppressor BRCA1 and inhibits BRCA1 activity. J Biol Chem 277: 7849–7856.
8. Sugihara TM, Bach I, Kioussi C, Rosenfeld MG, Andersen B (1998) Mouse Deformed epidermal autoregulatory factor 1 recruits a LIM domain factor, LMO-4, and CLIM coregulators. Proceedings of the National Academy of Sciences 95: 15418–15423.
9. Reed DE, Huang XM, Wohlschlegel JA, Levine MS, Senger K (2008) DEAF-1 regulates immunity gene expression in Drosophila. P Natl Acad Sci USA 105: 8351–8356.
10. Bottomley MJ, Collard MW, Huggenvik JI, Liu Z, Gibson TJ, et al. (2001) The SAND domain structure defines a novel DNA-binding fold in transcriptional regulation. Nat Struct Biol 8: 626–633.
11. Lutterbach B, Westendorf JJ, Linggi B, Patten A, Moniwa M, et al. (1998) ETO, a Target of t(8;21) in Acute Leukemia, Interacts with the N-CoR and mSin3 Corepressors. Mol Cell Biol 18: 7176–7184.
12. Gelmetti V, Zhang J, Fanelli M, Minucci S, Pelicci PG, et al. (1998) Aberrant Recruitment of the Nuclear Receptor Corepressor-Histone Deacetylase Complex by the Acute Myeloid Leukemia Fusion Partner ETO Mol Cell Biol 18: 7185–7191.
13. Liu Y, Chen W, Gaudet J, Cheney MD, Roudaia L, et al. (2007) Structural Basis for Recognition of SMRT/N-CoR by the MYND Domain and Its Contribution to AML1/ETO's Activity. Cancer Cell 11: 1–15.
14. Masselink H, Bernards R (2000) The adenovirus E1A binding protein BS69 is a corepressor of transcription through recruitment of N-CoR. Oncogene 19: 1538–1546.
15. Ansieau S, Leutz A (2002) The Conserved Mynd Domain of BS69 Binds Cellular and Oncoviral Proteins through a Common PXLXP Motif. JBC 277: 4906–4910.
16. Xu ST, Zhong C, Zhang TL, Ding JP (2011) Structure of human lysine methyltransferase Smyd2 reveals insights into the substrate divergence in Smyd proteins. Journal of Molecular Cell Biology 3: 293–300.
17. Xu ST, Wu J, Sun BF, Zhong C, Ding JP (2011) Structural and biochemical studies of human lysine methyltransferase Smyd3 reveal the important functional roles of its post-SET and TPR domains and the regulation of its activity by DNA binding. Nucleic Acids Research 39: 4438–4449.
18. Wang L, Li L, Zhang HL, Luo X, Dai JQ, et al. (2011) Structure of Human SMYD2 Protein Reveals the Basis of p53 Tumor Suppressor Methylation. Journal of Biological Chemistry 286: 38725–38737.
19. Sirinupong N, Brunzelle J, Ye J, Pirzada A, Nico L, et al. (2010) Crystal Structure of Cardiac-specific Histone Methyltransferase SmyD1 Reveals Unusual Active Site Architecture. Journal of Biological Chemistry 285.
20. Sirinupong N, Brunzelle J, Doko E, Yang Z (2011) Structural Insights into the Autoinhibition and Posttranslational Activation of Histone Methyltransferase SmyD3. Journal of Molecular Biology 406: 149–159.
21. Jiang YY, Sirinupong N, Brunzelle J, Yang Z (2011) Crystal Structures of Histone and p53 Methyltransferase SmyD2 Reveal a Conformational Flexibility of the Autoinhibitory C-Terminal Domain. Plos One 6.
22. Foreman KW, Brown M, Park F, Emtage S, Harriss J, et al. (2011) Structural and Functional Profiling of the Human Histone Methyltransferase SMYD3. PLoS ONE 6: e22290.
23. Ferguson AD, Larsen NA, Howard T, Pollard H, Green I, et al. (2011) Structural Basis of Substrate Methylation and Inhibition of SMYD2. Structure 19: 1262–1273.
24. Borden KLB, Freemont PS (1996) The RING finger domain: A recent example of a sequence-structure family Curr Opin Struct Biol 6: 395–401.
25. Spadaccini R, Perrin H, Bottomley MJ, Ansieau S, Sattler M (2006) Structure and functional analysis of the MYND domain. J Mol Biol 358: 498–508.
26. Spadaccini R, Perrin H, Bottomley MJ, Ansieau S, Sattler M (2008) Retraction notice to "Structure and functional analysis of the MYND domain" [J. Mol. Biol. (2006) 358, 498–508]. J Mol Biol 376: 1523.
27. Pelton JG, Torchia DA, Meadow ND, Roseman S (1993) Tautomeric States of the Active-Site Histidines of Phosphorylated and Unphosphorylated III(GLC), A Signal-Transducing Protein from Escherichia-coli, Using 2-Dimensional Heteronuclear NMR Techniques. Protein Sci 2: 543–558.
28. Barlow PN, Luisi B, Milner A, Elliott M, Everett R (1994) Structure of the C3HC4 Domain by H-1-Nuclear Magnetic-Resonance Spectroscopy - A New Structural Class of ZINC-Finger. J Mol Biol 237: 201–211.
29. Sattler M, Schleucher J, Griesinger C (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients Prog Nucl Magn Reson Spectrosc 34: 93–158.
30. Shen Y, Delaglio F, Cornelisscu G, Bax A (2009) TALOS+: A hybrid method for predicting protein backbone torsion angles from NMR chemical shifts J Biomol NMR 44: 213–223.
31. Güntert P (2009) Automated structure determination from NMR Spectra. Eur Biophys J 38: 129–143.
32. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, et al. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54: 905–921.
33. Linge JP, Habeck M, Rieping W, Nilges M (2003) ARIA: automated NOE assignment and NMR structure calculation. Bioinformatics 19: 315–316.
34. Linge JP, Williams MA, Spronk CA, Bonvin AM, Nilges M (2003) Refinement of protein structures in explicit solvent. Proteins 50: 496–506.
35. Schwabe JWR, Klug A (1994) Zinc mining forr protein domains. Structural Biology 1: 345–348.
36. Lipari G, Szabo A (1982) Model-Free Approach to the Interpretation of Nuclear Magnetic Resonance Relaxation In Macromolecules 1. Theory and Range of Validity. J Am Chem Soc 104: 4546–4559.
37. Lipari G, Szabo A (1982) Model-Free Approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. 2. Analysis of experimental Results. J Am Chem Soc 104: 4559–4570.
38. Dosset P, Marion D, Blackledge M (2001) TENSOR2. Copyright Institut de Biologie Structurale J.P. Ebel, CEA-CNRS.
39. Freemont PS, Hanson IM, Trowsdale J (1991) A Novel Cysteine-Rich Sequence Motif. Cell 64: 483–484.
40. Lutterbach B, Sun D, Schuetz J, Hiebert SW (1998) The MYND Motif Is Required for Repression of Basal Transcription from the Multidrug Resistance 1 Promoter by the t(8;21) Fusion Protein. Mol Cell Biol 18: 3604–3611.
41. Ansieau S, Leutz A (2002) The conserved Mynd domain of BS69 binds cellular and oncoviral proteins through a common PXLXP motif. J Biol Chem 277: 4906–4910.
42. Marti-Renom M, Stuart A, Fiser A, Sanchez R, Melo F, et al. (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomolec Struct 29: 291–325.
43. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, et al. (2002) Comparative Protein Structure Modeling Using Modeller: John Wiley & Sons, Inc.
44. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234: 779–815.
45. Fiser A, Do RKG, Šali A (2000) Modeling of loops in protein structures. Protein Science 9: 1753–1773.
46. Ansieau S, Sergeant A (2003) BS69 and RACK7, a potential novel class of tumor suppressor genes. Pathol Biol 51: 397–399.
47. Savage CR, Hash JH, Cohen S (1973) Epidermal Growth Factor. Journal of Biological Chemistry 248: 7669–7672.
48. Lausen J, Cho S, Liu S, Wener MH (2004) The nuclear receptor co-repressor (N-CoR) utilizes repression domains I and III for interaction and co-repression with ETO. J Biol Chem 279: 49281–49288.
49. Sims RJ, Weihe EK, Zhu L, O'Malley S, Harriss JV, et al. (2002) m-Bop, a Repressor Protein Essential for Cardiogenesis, Interacts with skNAC, a Heart- and Muscle-specific Transcription Factor. Journal of Biological Chemistry 277: 26524–26529.
50. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, et al. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31: 3784–3788.
51. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, et al. (1995) NMRPipe: a Multidimensional Spectral Processing System Based on UNIX Pipes. J Biomol NMR 6: 277–293.
52. Goddard TD, Kneller DG (2008) SPARKY 3.

53. Cordier F, Dingley AJ, Grzesiek S (1999) A doublet-separated sensitivity-enhanced HSQC for the determination of scalar and dipolar one-bond J-couplings J Biomol NMR 13: 175–180.
54. Yang D, Tolman JR, Goto NK, Kay LE (1998) An HNCO-based pulse scheme for the measurement of C-13(alpha)-H-1(alpha) one-bond dipolar couplings in N-15, C-13 labeled proteins J Biomol NMR 12: 325–332.
55. Yang D, Venters RA, Mueller GA, Choy WY, Kay LE (1999) TROSY-based HNCO pulse sequences for the measurement of (HN)-H-1-N-15, N-15-(CO)-C-13, (HN)-H-1-(CO)-C-13, (CO)-C-13-C-13(alpha) and (HN)-H-1-C-13(alpha) dipolar couplings in N-15, C-13, H-2-labeled proteins J Biomol NMR 14: 333–343.
56. Kay LE, Torchia DA, Bax A (1989) Backbone Dynamics of Proteins as Studied by N-15 Inverse Detected Heteronuclear NMR-Spectroscopy - Application to Staphylococcal Nuclease. Biochemistry 28: 8972–8979.
57. Johnson BA, Blevins RA (1994) NMR View: A computer program for the visualization and analysis of NMR data. Journal of Biomolecular NMR 4: 603–614.
58. Clore GM, Gronenborn AM, Bax A (1998) A Robust Method for Determining the Magnitude of the Fully Asymmetric Alignment Tensor of Oriented Macromolecules in the Absence of Structural Information. J Magn Reson 133: 216–221.
59. Nilges M (1995) Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. J Mol Biol 245: 645–660.
60. Nilges M, O'Donoghue S (1998) Ambiguous NOEs and automated NOE assignment. Prog Nucl Magn Reson Spectrosc 32: 107–139.
61. Linge JP, Nilges M (1999) Influence of non-bonded parameters on the quality of NMR structures: a new force-field for NMR structure calculation. J Biomol NMR 13: 51–59.
62. Linge JP (2001) New methods for automated NOE assignment and NMR structure calculation.
63. Linge JP, O'Donoghue S, Nilges M (2001) Assigning Ambiguous NOEs with ARIA. Meth Enzymol 339: 71–90.
64. Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8: 477–486.
65. Velasco G, Grkovic S, Ansieau S (2006) New Insights into BS69 Functions. Journal of Biological Chemistry 281: 16546–16550.
66. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: Application to microtubules and the ribosome. Proceedings of the National Academy of Sciences 98: 10037–10041.

THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 289, NO. 41, pp. 28640–28650, October 10, 2014
© 2014 by The American Society for Biochemistry and Molecular Biology, Inc. Published in the U.S.A.

A Novel Protein-Protein Interaction in the RES (REtention and Splicing) Complex*

Received for publication, June 26, 2014, and in revised form, August 12, 2014 Published, JBC Papers in Press, August 26, 2014, DOI 10.1074/jbc.M114.592311

Konstantinos Tripsianes[‡1], Anders Friberg[§¶2], Charlotte Barrandon[‖3], Mark Brooks[4], Herman van Tilbeurgh[**], Bertrand Seraphin[‖‡‡], and Michael Sattler[§¶5]**

From the [‡]Central European Institute of Technology (CEITEC), Masaryk University, Kamenice 5, 62500 Brno, Czech Republic, the [§]Institute of Structural Biology, Helmholtz Zentrum München, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany, the [¶]Center for Integrated Protein Science Munich and Chair of Biomolecular NMR, TU München, Lichtenbergstr. 4, 85747 Garching, Germany, the [‖]Centre de Génétique Moléculaire, CNRS, Avenue de la Terrasse, 91198 Gif sur Yvette, France, the [**]University Paris-Sud, Institut de Biochimie et de Biophysique Moléculaire et Cellulaire, UMR8619, F-91405 Orsay, France, and the [‡‡]Equipe Labellisée La Ligue, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), Centre National de Recherche Scientifique (CNRS) UMR 7104, Institut National de Santé et de Recherche Médicale (INSERM) U964, Université de Strasbourg, 67404 Illkirch, France

Background: The heterotrimeric retention and splicing (RES) complex is conserved from yeast to man.
Results: Structural and biochemical analyses show an extended binding interface between the Snu17p RNA recognition motif (RRM) and Bud13p of the RES complex.
Conclusion: The Snu17p RRM-Bud13p interaction represents a novel RRM-protein interface.
Significance: The Snu17p-Bud13p complex may provide a platform for additional protein and RNA interactions in the RES complex.

The retention and splicing (RES) complex is a conserved spliceosome-associated module that was shown to enhance splicing of a subset of transcripts and promote the nuclear retention of unspliced pre-mRNAs in yeast. The heterotrimeric RES complex is organized around the Snu17p protein that binds to both the Bud13p and Pml1p subunits. Snu17p exhibits an RRM domain that resembles a U2AF homology motif (UHM) and Bud13p harbors a Trp residue reminiscent of an UHM-ligand motif (ULM). It has therefore been proposed that the interaction between Snu17p and Bud13p resembles canonical UHM-ULM complexes. Here, we have used biochemical and NMR structural analysis to characterize the structure of the yeast Snu17p-Bud13p complex. Unlike known UHMs that sequester the Trp residue of the ULM ligand in a hydrophobic pocket, Snu17p and Bud13p utilize a large interaction surface formed around the two helices of the Snu17p domain. In total 18 residues of the Bud13p ligand wrap around the Snu17p helical surface in an U-turn-like arrangement. The invariant Trp^{232} in Bud13p is located in the center of the turn, and contacts surface residues of Snu17p. The structural data are supported by mutational analysis and indicate that Snu17p provides an extended binding surface with Bud13p that is notably distinct from canonical UHM-ULM interactions. Our data highlight structural diversity in RRM-protein interactions, analogous to the one seen for nucleic acid interactions.

RNA splicing is one of the fundamental processes in constitutive and regulated gene expression in eukaryotes. During splicing, introns present in primary transcripts (pre-mRNAs) are removed and exons are ligated to produce translationally competent mRNAs. The reaction is mediated by the spliceosome, a highly dynamic machinery of five small nuclear ribonucleoprotein particles (snRNPs)[6] and a large number of non-snRNP interacting proteins (1). In the course of the splicing reaction snRNPs assemble stepwise and undergo substantial conformational rearrangements that allows processing of a wide variety of transcripts. In addition, non-snRNP proteins associate and exchange at different stages of the splicing process, allowing for further splicing regulation or coupling of splicing to other cellular processes (2).

* This work was supported, in whole or in part, by the European Social Fund and state budget of the Czech Republic Grant CZ.1.07/2.3.00/20.0042 (to K. T.), the Centre National pour la Recherche Scientifique (to B. S.), Grant ANR-10-LABX-0030-INRT managed under the program Investissements d'Avenir ANR-10-IDEX-0002-02 (to B. S.), the European Union Sixth Framework program "3D-Repertoire" Grant LSHG-CT-2005-512028 (to B. S. and M. S.) and EURASNET Grant LSHG-CT-2005-518238 (to B. S.), Ligue Contre le Cancer (Equipe Labellisée 2014) (to B. S.), the CERBM-IGBMC (to B. S.), and Deutsche Forschungsgemeinschaft Grant SFB1035 (to M. S.).
⌘ *Author's Choice*—Final version full access.
The atomic coordinates and structure factors (code 4UQT) have been deposited in the Protein Data Bank (http://wwpdb.org/).
The chemical shifts are available at the Biological Magnetic Resonance Bank (BMRB) with accession number 25047.
[1] To whom correspondence may be addressed: Central European Institute of Technology, Masaryk University, Kamenice 5, 62500 Brno, Czech Republic. Tel.: 420-549-49-6607; Fax: 420-549-49-2556; E-mail: kostas.tripsianes@ceitec.muni.cz.
[2] Supported by Elitenetzwerk Bayern. Present address: Bayer HealthCare Pharmaceuticals, Lead Discovery, Müllerstr. 178, 13353 Berlin, Germany.
[3] Present address: Promega France, Parc d'Activités des Verrières, 24 Chemin des Verrières, 69260 Charbonnières, France.
[4] Present address: Evotec (UK) Ltd., 114 Milton Park, Abingdon OX14 4SA, United Kingdom.
[5] To whom correspondence may be addressed: Institute of Structural Biology, Helmholtz Zentrum München, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany. Tel.: 49-89-3187-3800; Fax: 49-89-289-13869; E-mail: sattler@helmholtz-muenchen.de.

[6] The abbreviations used are: snRNP, small nuclear ribonucleoprotein particle; RRM, RNA recognition motif; ITC, isothermal titration calorimetry; CSP, chemical shift perturbation; Ni-NTA, nickel-nitrilotriacetic acid; HSQC, heteronuclear single quantum coherence.

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

A non-snRNP complex has been recently identified in *Saccharomyces cerevisiae* that enhances pre-mRNA splicing and prevents leakage of unspliced pre-mRNAs from the nucleus (3). The yeast pre-mRNA retention and splicing complex, called the RES complex, consists of Snu17p, Bud13p, and Pml1p proteins. Snu17p acts as a central platform, which independently binds Bud13p and Pml1p (4–6). The RES subunits and their human orthologs have been shown to associate transiently with the spliceosome prior to the first catalytic step that leads to intron excision (7, 8). Yeast strains carrying deletions of *snu17*, *bud13*, or *pml1* are viable but exhibit slow growth, a phenotype exacerbated at high temperatures. *In vitro* and *in vivo* splicing assays have demonstrated that the RES complex enhances splicing, especially of those introns that contain poor consensus sequence at the 5′ splice site (3, 9), including alternatively spliced genes (10). Inactivation of RES subunits also induces leakage of pre-mRNAs from the nucleus to the cytoplasm, indicating that the RES complex has an important role in nuclear retention of unspliced transcripts (3). Finally, the RES complex was recently shown to facilitate the splicing of introns with short 5′ splice site-branch point distances (11) and to interact genetically with the enzyme mediating trimethylation of the U snRNAs (12).

Of the three RES subunits, Snu17p contains an RNA recognition motif (RRM) domain (Fig. 1). The RRM motif is the most abundant RNA-binding domain in higher vertebrates and found in many spliceosomal proteins. RRMs usually exhibit a compact-fold comprised of four- or five-stranded β-sheet that represents the RNA binding surface and two α helices packed against the β-sheet. Sequence conservation between various RRMs is low except for three aromatic side chains belonging to the two signature sequences RNP1 and RNP2, which are responsible for RNA binding (Fig. 1, *a* and *c*) (13, 14). The UHM (U2AF homology motif) domain is a variant of the RRM domain and has been shown to mediate protein-protein interactions via the α-helical face of the RRM-fold (15). UHM domains often lack aromatic residues in the RNP signature sequences and thus do not or only poorly bind to RNA. An Arg-*X*-Phe motif in UHM domain is involved in the recognition of a conserved tryptophan residue in the UHM ligand motifs (ULMs). Structural studies have demonstrated a common mode of ligand recognition by different UHM domains (15–19). The invariant Trp residue in the ULM (Fig. 1*b*) inserts into a tight hydrophobic pocket formed by the two helices and the Arg-*X*-Phe motif of the UHM domain (Fig. 1*d*), whereas positively charged residues preceding the conserved Trp in the ULM motif interact with negatively charged residues on the UHM domain.

Previous biochemical studies have indicated that the RRM domain of Snu17p is sufficient for binding Bud13p. Although some unconventional features of the Snu17p domain were noticed, such as the absence of the Arg-*X*-Phe motif (Fig. 1*a*), it has been proposed that the Snu17p-Bud13p interaction follows the structural characteristics of UHM-ULM interactions (4, 6, 20). To better understand how the RES complex is organized to perform its biological functions we have determined the binary complex structure between Snu17p and Bud13p. Isothermal titration calorimetry (ITC) together with NMR binding experiments allowed the identification of the minimal length of the Bud13p ligand. The NMR-derived solution structure shows that Snu17p and Bud13p do not interact through an UHM-ULM type of recognition. Rather, the interaction involves 18 residues of Bud13p that contact the helical surface of the Snu17p RRM domain. Notably, the invariant Trp of Bud13p is not buried in a hydrophobic cavity, distinct from canonical UHM-ULM interactions. The structural data are supported by mutational analysis of the interacting partners testing for their ability to form the RES complex when co-expressed in *Escherichia coli* or to restore the growth phenotype of Δ*snu17* or Δ*bud13* strains.

EXPERIMENTAL PROCEDURES

Cloning, Protein Expression, and Purification—Snu17p proteins (residues 1–113 or 25–113) and Bud13p proteins (residues 201–266 or 222–256) were cloned into a modified pET-24d vector using standard protocols. The fusion proteins expressed from these vectors comprise a His_6 tag followed by a GST fusion domain and a tobacco etch virus proteolytic cleavage site. Proteins were expressed in *Escherichia coli* BL21(DE3) pLysS (Novagen) using kanamycin for selection. A 10-ml Luria broth (LB) pre-culture was inoculated with a single colony from a transformation plate. The pre-culture was used to start larger 1-liter cultures, containing LB or M9 minimal medium for labeling with ^{15}N or $^{15}N/^{13}C$. Upon reaching optical density of 0.6 cultures were induced with 0.5 mM isopropyl β-D-1-thiogalactopyranoside for 4 h at 37 °C.

Recombinant proteins were purified by sonicating the harvested cell pellet in 25 ml of lysis buffer (20 mM Tris, pH 7.5, 300 mM NaCl, 10 mM imidazole, 1 mM DTT, and 0.02% NaN_3), also including protease inhibitors, RNase, lysozyme, and 0.2% IGEPAL. After high speed centrifugation (20,000 × *g*, 30 min) and filtering, each supernatant was applied to Ni-NTA-agarose resin (Qiagen). Several rounds of washing were performed with: lysis buffer including 0.2% IGEPAL, lysis buffer, lysis buffer with high salt concentration (1 M NaCl), and lysis buffer with high imidazole concentration (30 mM imidazole). Finally, each protein was eluted by applying 10 ml of a buffer containing 20 mM Tris, pH 7.5, 300 mM NaCl, 330 mM imidazole, 1 mM DTT, and 0.02% NaN_3. Tobacco etch virus protease was added to each sample and incubated overnight at 4 °C. To remove the cleaved GST tag each sample was passed over a second Ni-NTA column. A second purification step included size exclusion chromatography (HiLoad, Superdex 75 16/60, GE Healthcare). In this step, proteins were buffer exchanged into the NMR buffer (20 mM sodium phosphate, pH 6.3, 25 mM NaCl, and 2 mM fresh DTT or Tris(2-carboxyethyl)phosphine, 0.02% NaN_3). Stoichiometric hybrid-labeled samples were prepared by mixing excess Bud13p with Snu17p and performing another step of size exclusion chromatography (HiLoad, Superdex 75 16/60, GE Healthcare). Two shorter peptides of Bud13p (residues 221–242 or residues 225–238) were purchased from Peptide Specialty Laboratories (Heidelberg, Germany), and extensively dialyzed before binding experiments.

An operon system encoding His_6-Snu17p, Bud13p, and Pml1p was described earlier (4). Snu17p and Bud13p variants were introduced into this vector by standard mutagenesis and cloning procedures. Protein expression, complex mini-purifi-

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

cations, and denaturing gel analyses of complex composition were performed as previously described (4).

NMR Spectroscopy and Structure Calculation—NMR spectra were recorded at 298 K on an AVIII800 Bruker NMR spectrometer equipped with a cryogenic triple resonance gradient probe. Spectra were processed using NMRPipe (21). Protein backbone and side chain assignments were obtained from standard triple resonance experiments (22). NOE distance restraints were obtained from ^{15}N- and ^{13}C-edited three-dimensional NOESY spectra (mixing time 70 ms) where only Snu17p or Bud13p proteins were labeled. Intermolecular NOE restraints were derived from three-dimensional ^{15}N,^{13}C-filtered, ^{13}C-edited NOESY spectra (mixing time 70 ms) (22, 23). Stereospecific assignments of pro-chiral methyl groups of valine and leucine were obtained using 10% ^{13}C-labeled protein samples (24).

Structure calculations with torsion angle dynamics were performed by using the program CYANA 3.0 (25). For every assigned peak, the distances between the protons (for a range between 2.4 and 6 Å) were calibrated using the calibration algorithm of CYANA. The set of NOE distance restraints together with φ and ψ backbone dihedral angle restraints derived from TALOS+ (26) based on the chemical shifts were used for water refinement (27) using CNS (28). The structure was validated using iCing. Molecular images were generated with PyMol (Schrödinger).

Isothermal Titration Calorimetry—Calorimetric measurements were performed with a VP-ITC (MicroCal Inc.) instrument at 25 °C using samples in NMR buffer. The concentration of Snu17p (residues 1–113) in the cell (1440 μl) varied between 30 and 70 μM, whereas the syringe (300 μl) contained Bud13p ligands at concentrations ranging from 0.6 to 1.1 mM. Normally 50 injections were done over a period of 3–4 h. Bud13p ligands were also injected into buffer and the heat of dilution control was subtracted from the titration. Data fitting was done in Origin v7 using one- or two-step binding models supplied by the manufacturer. The dissociation constant values are given with the fitting error.

In Vivo Analyses—The wild type *snu17* and *bud13* genes were cloned in the pRS425 plasmid backbone. The TAP tag coding sequence (29) was introduced to produce C-terminal fusions, allowing the monitoring of protein expression levels. *snu17* and *bud13* variants were introduced into the resulting vectors by standard mutagenesis and cloning procedures. These plasmids were transformed, respectively, into Δ*snu17* or Δ*bud13* strains (3, 4). Transformants were selected and complementation was assayed by spotting serial dilutions of the strain on selective plates and incubating at the indicated temperatures. Expression of wild type and mutant proteins was monitored by extracting total proteins and monitoring the levels of the TAP-tagged fusions by Western blotting.

RESULTS

RES Complex Assembly and Function—Previous biochemical analyses have shown that Snu17p binds Bud13p directly (4, 6). Based on a mutational analysis of the invariant Bud13p Trp^{232} residue (6), the interaction of Bud13p with the Snu17p RRM has been proposed to resemble a typical UHM-ULM interaction. Another study, however, showed that the W232A mutation did not abolish its interaction with Snu17p (4), in contrast to what has been reported so far for UHM-ULM interactions (18). Moreover, it had been noticed that Snu17p lacks the UHM characteristic Arg-*X*-Phe motif and also contains the RRM-typical consensus sequence motifs (RNP1 Phe^{74}, Tyr^{76}, and RNP2 Tyr^{34}) that mediate RNA interactions in canonical RRM domains (Fig. 1, *a* and *c*).

To characterize the Snu17p-Bud13p interaction *in vivo* we performed mutational analysis in *S. cerevisiae*. We first tested the impact of mutations on complex formation *in vitro* by co-expressing His$_6$-Snu17p, Bud13p, and Pml1p encoded by an operon in *E. coli* and assessing complex assembly by purification on Ni-NTA (4). A double or triple mutant around the conserved Trp residue (R231D,W232A or R231D,W232A,D233N) of Bud13p abolished Bud13p incorporation in the RES complex (Fig. 2*a*). To assay the *in vivo* function, a yeast strain lacking the *bud13* gene (Δ*bud13*) was transformed with a vector carrying either the wild type Bud13p tagged with the TAP tag, or its double or triple mutant derivatives. Cells carrying the vector without insert were used as a negative control. At 30 °C, the growth of both mutants was only mildly affected (slightly smaller colony sizes than wild type cells, Fig. 2*b*). Consistent with previous observations deletion of *bud13* resulted in a strong phenotype at 37 °C (3). In this condition, both mutants were barely different from cells transformed with the vector indicating that they were not functional. Transformation of the same *bud13* mutants in a wild type strain did not result in slow growth demonstrating that they were not functional rather than acting as dominant negatives. Moreover, growth deficiency was not due to a decreased expression of the mutant proteins (detected by Western blot analyses of the TAP tag) (Fig. 2*c*) excluding that the mutations acted simply by destabilizing the proteins. These data indicate that Trp^{232} of Bud13p indeed is important for the interaction with Snu17p but that flanking residues also contribute to the binding.

Snu17p mutants were designed based on sequence homology with UHM domains that utilize a hydrophobic pocket between the two helices of the UHM domain to accommodate the invariant Trp of the ULM ligand. We targeted Snu17p residues corresponding to the Arg-*X*-Phe motif (^{95}FKI^{97}). A triple mutant (F95A,K96D,I97A) did not disrupt RES complex formation as assayed by co-expression of the proteins in *E. coli* followed by Ni-NTA purification (Fig. 2*a*). We also tested the ability of this mutant to rescue the growth phenotype of Δ*snu17* yeast cells. The mutant grew well at all temperatures tested (Fig. 2*d*) both in the mutant and wild type strain backgrounds. Western blotting confirmed that the mutant was well expressed (Fig. 2*e*). These results suggest that Snu17p may use a different interaction surface than canonical UHM proteins to bind the Bud13p ligand.

Biophysical Analysis of Snu17p-Bud13p Complex—To characterize the unexpected molecular features of the Snu17p-Bud13p interaction we proceeded with biophysical and structural characterization of the binary complex (Fig. 3). We first used a construct comprising the Snu17p RRM and an N-terminal unstructured (see below) extension (Snu17p, residues 1–113) as this domain was shown to be sufficient for the interaction with Bud13p (4, 6). We tested binding to a series of

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

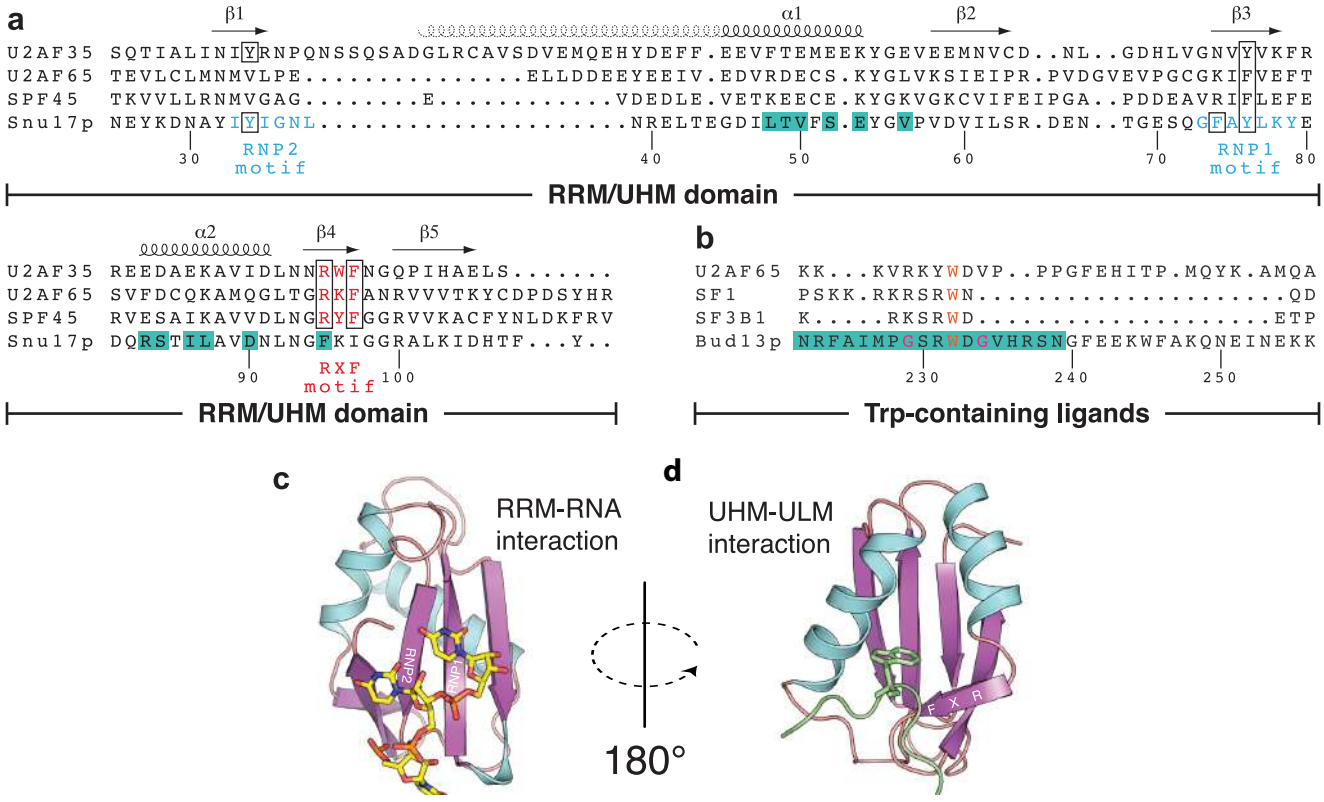


FIGURE 1. **Sequence and structural features of RRM and UHM domains.** *a*, structure-based sequence alignment of UHM/RRM domains. Secondary elements are indicated on the *top* and residue numbers of Snu17p are given at the *bottom*. Residues in the RNP motifs of Snu17p and Arg-*X*-Phe motifs of UHM domains are colored *cyan* and *red*, respectively. Conserved residues in the motifs are highlighted by *black boxes*. *b*, sequence alignment of ULM peptide sequences known to bind UHM domains with the Bud13p ligand (residue numbers of Bud13p indicated at the bottom). The invariant tryptophan of ULM ligands and the two glycine residues that are unique to Bud13p are colored *orange* and *magenta*, respectively. Residues that form the binding interface between Snu17p and Bud13p are highlighted with a *lime green background* in *a* and *b*, respectively. *c* and *d*, structures of a canonical RRM domain (PDB code 2G4B) in complex with RNA (*c*) and a representative UHM domain (PDB 2PEH) bound to a ULM peptide (*d*). RNP motifs and the UHM-specific RXF sequence are annotated in *β* strands.

Bud13p fragments to identify the minimum binding region in Bud13p (Fig. 3, *a* and *b*). NMR titrations confirmed binding to a Bud13p peptide encompassing residues 201–266 as has been shown previously (20). Addition of this peptide to recombinant ^{15}N-labeled Snu17p induced extensive chemical shift perturbations (CSPs) in ^{1}H,^{15}N correlation spectra but also led to precipitation of the sample. Similar CSPs were observed with a shorter Bud13p peptide (residues 222–256) but proteins remained in solution (Fig. 3*b*). An even shorter peptide (residues 221–242) affected the same Snu17p residues but to a lesser extent compared with the longer ones. Finally, a very short Bud13p peptide (residues 225–238) hardly induced any NMR spectral changes for Snu17p, even though this peptide has the typical length of ULM ligands and harbors the invariant Trp (Fig. 3*b*). Thus, the NMR data indicate that recognition of an ULM-like motif with its central Trp residue is not sufficient for Bud13p-Snu17p complex formation. This further demonstrates a different recognition mode involving additional residues present in the longer Bud13p peptides.

The Snu17p-Bud13p interaction was further characterized by ITC to quantitate the binding affinity and characterize the thermodynamic features of the interaction. The binding isotherms of the two longer peptides show low nanomolar binding affinity. The calorimetric analysis also suggests a second interaction at large excess of Bud13p, however, with a lower affinity with binding constants separated by 2 orders of magnitude (Fig. 3*a*), potentially reflecting an additional nonspecific interaction. In any case, the ITC data corroborate NMR findings that the additional residues present in the very long Bud13p peptide do not contribute to Snu17p binding when compared with the Bud13p-(222–256) peptide. The low nanomolar binding affinities are in very good agreement with the thermodynamic binding parameters of the full-length Snu17p-Bud13p complex (6). The ITC titration of Snu17p with the shorter Bud13p peptide (residues 221–242) could be fitted to a single binding site with a dissociation constant 50 times weaker compared with the longer ligands (Fig. 3*a*). The lower affinity is in accordance with quantitative information obtained from NMR chemical shift titration analysis for the Bud13p ligands (data not shown). For the shortest Bud13p peptide, the signal to noise ratio of the ITC signal was too small to make a precise determination of its binding constant, indicating that it binds very weakly consistent with the NMR titrations (Fig. 3, *a* and *b*). Based on NMR and ITC data we conclude that residues 222–256 of Bud13p encompass the minimum interacting region with Snu17p.

Next, we assigned the backbone chemical shifts of Snu17p (residues 1–113) when free and in a stoichiometric complex with unlabeled Bud13p (residues 222–256). Chemical shift analysis shows that Snu17p contains the expected secondary structure elements of an RRM domain in both free and bound forms (Fig. 3*c*), *i.e.* the secondary structure is preserved upon Bud13p binding. The first 25 residues of Snu17p do not exhibit

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

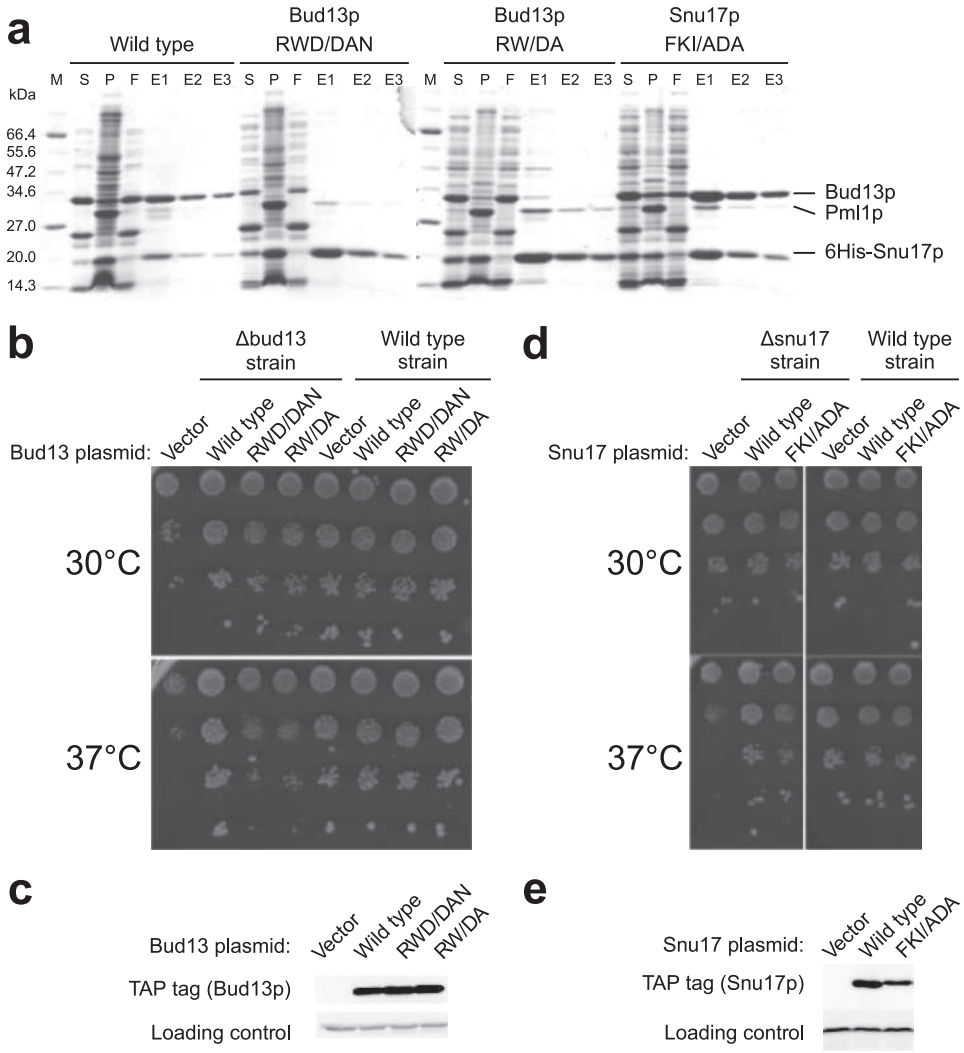


FIGURE 2. **Mutational analyses of the Bud13p-Snu17p interaction.** *a,* interactions between Snu17p, Bud13p, and Pml1p were assayed by one-step Ni-NTA purification of wild type or mutant proteins co-expressed in *E. coli*. Supernatant (*S*) and pellet (*P*) fractions from bacterial lysates as well as flow-through (*F*) and elution (E1, E2, and E3) fractions from the Ni-NTA purifications are shown after Coomassie staining of the gels. *M* indicates the molecular mass marker with size indicated on the *left* in kDa. Note that Pml1p is less well stained than Snu17p and Bud13p. *b,* complementation of Δ*bud13* strains with an empty vector, or plasmids expressing wild type or mutant Bud13p was monitored through the growth of serial dilutions of transformants on selective plates at the indicated temperatures. *c*, Western blot showing the level of wild type and mutant Bud13p is detected through the fused TAP tag. *d,* complementation of a Δ*snu17* strain was performed as for *bud13* in *b*. *e,* the levels of wild type and mutant Snu17p were monitored by detection of the TAP tag by Western blotting.

any secondary structure and $\{^1H\}$-^{15}N heteronuclear NOE data show that the N terminus preceding the RRM domain is highly flexible (Fig. 3*c*). The residues showing large chemical shift changes, induced by Bud13p binding, map on the two helices plus adjacent loop regions of the Snu17p RRM-fold indicating a large interaction surface (Fig. 3*c*). These CSPs are distinct from what would be expected for a canonical UHM-ULM interaction, demonstrating that Snu17p and Bud13p use a different binding interface. Virtually identical NMR CSPs induced by Bud13p binding to a shorter Snu17p protein (residues 25–113) demonstrated that the flexible Snu17p N terminus does not participate in the interaction (data not shown).

The Snu17p-Bud13p interaction was also characterized by monitoring NMR signals of ^{15}N-labeled Bud13p (residues 222–256) in complex with unlabeled Snu17p. Binding to Snu17p induces extensive chemical shift changes in Bud13p spectra (Fig. 4*a*). Secondary chemical shift analysis indicates the presence of an α-helical segment at the C terminus of the bound Bud13p peptide (Fig. 4*b*). Steady state $\{^1H\}$-^{15}N heteronuclear NOE data show that a large portion of the peptide is rigid in the complex suggesting that the region comprising Bud13p residues 222–246 contributes to the Snu17p binding interface directly or become less flexible due to complex formation (Fig. 4*b*). Interestingly, the C-terminal helical region remains partially flexible.

Structure of the Snu17p-Bud13p Complex—We then determined the solution structure of the Snu17p-Bud13p complex (comprising Snu17p residues 25–113 and Bud13p residues 222–256) by heteronuclear triple-resonance NMR spectroscopy using subunit selectively isotope-labeled samples. The two proteins were prepared separately and co-purified by size exclusion chromatography after mixing an excess of Bud13p with Snu17p. In this way we were able to selectively label each of the proteins in the complex and determine 129 intermolecular distance restraints, derived from ^{13}C- and ^{15}N-edited NOESY HSQC spectra as well as $^{13}C/^{15}N$-filtered and ^{13}C-edited

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

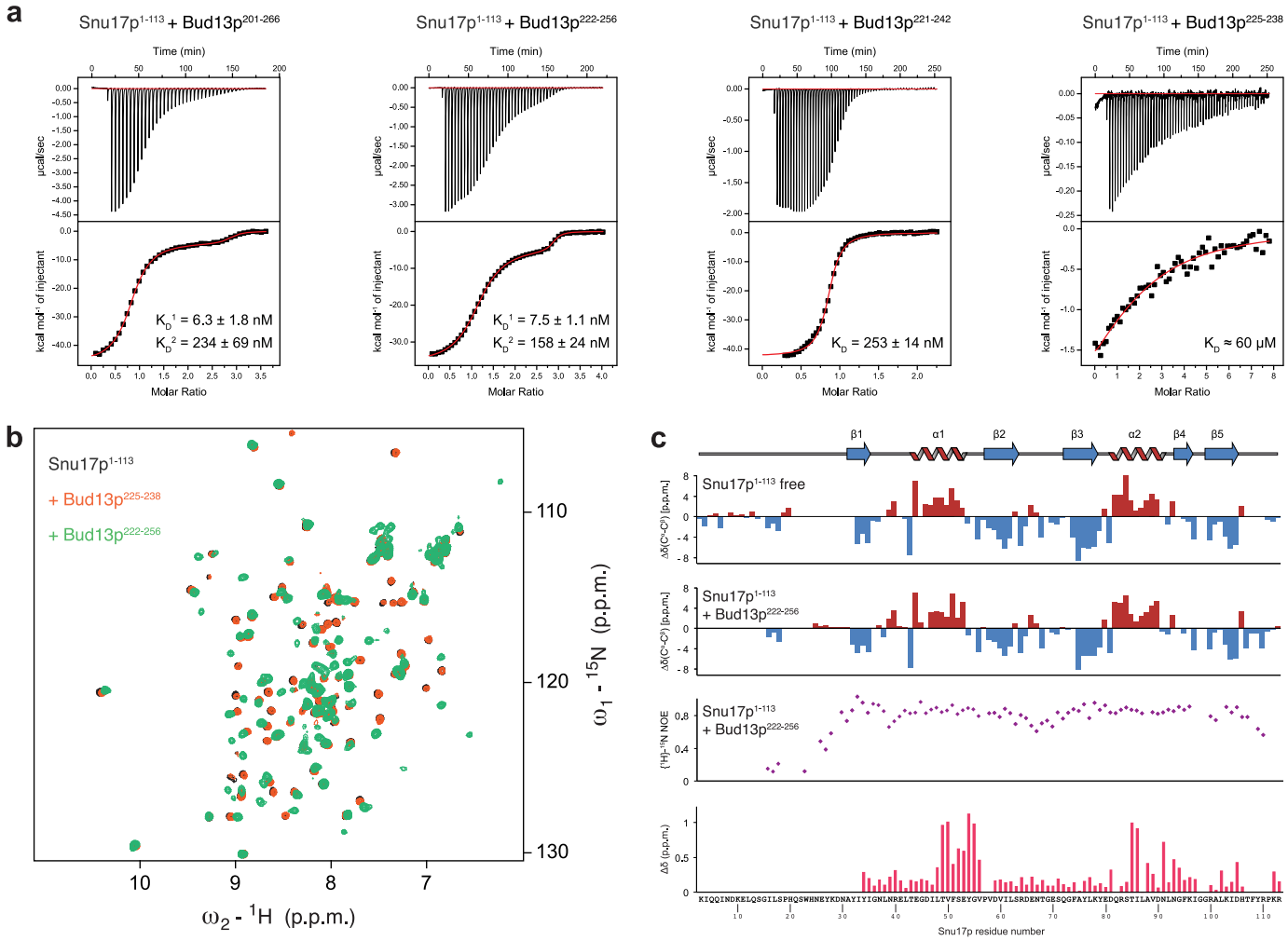


FIGURE 3. **ITC and NMR analysis of Snu17p binding to Bud13p.** *a,* four different Bud13p peptides containing the invariant Trp residue were tested by ITC titrations for Snu17p binding. Binding curves for a 66-mer, a 35-mer used in the accompanying structural studies, a 22-mer, and a 14-mer Bud13p peptide, are depicted from *left to right*. Dissociation constants obtained by fitting the isotherms after subtracting the heat of dilution are listed. *b,* overlay of $^{1}H,^{15}N$ HSQC spectra of free Snu17p (residues 1–113, *black*) and after titration with the shortest Bud13p ligand (residues 225–238, *orange*) or with the optimal Bud13p peptide (residues 222–256, *green*). *c,* secondary ^{13}C chemical shifts $\Delta\delta(^{13}C^{\alpha})$-$\Delta\delta(^{13}C^{\beta})$ for Snu17p free and bound to Bud13p-(222–256), $\{^{1}H\}$-^{15}N heteronuclear NOE of the Snu17p-Bud13p complex, and chemical shift perturbations of Snu17p amide signals (recorded on a sample of ^{15}N-labeled Snu17 residues 25–113) upon binding to Bud13p-(222–256) are plotted *versus* the amino acid sequence of Snu17p. Secondary structural elements of Snu17p are indicated on *top*.

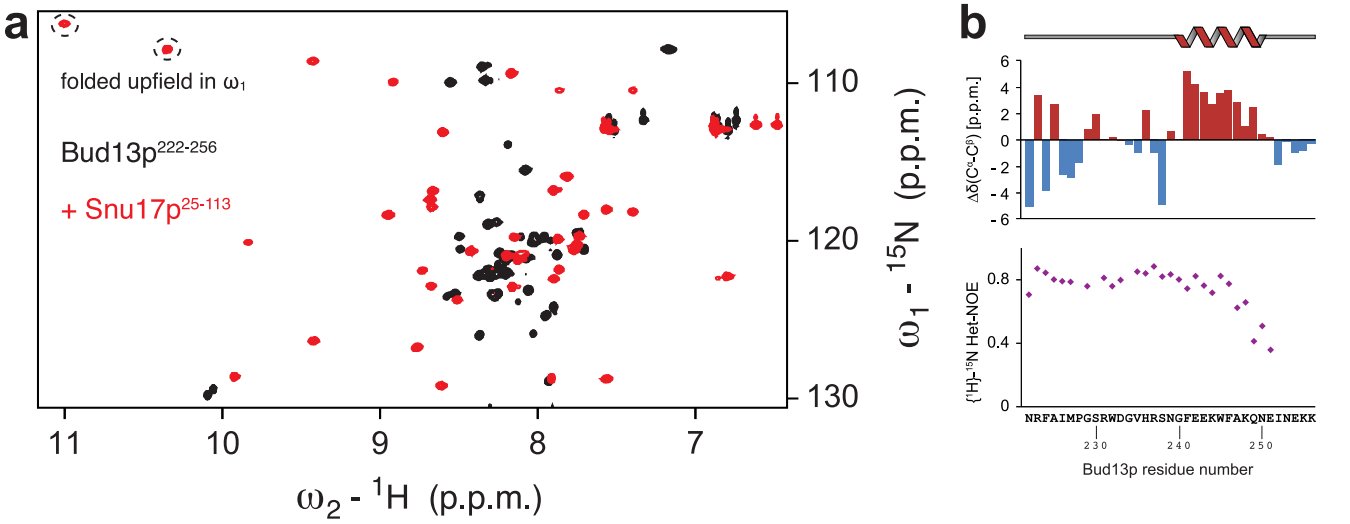


FIGURE 4. **NMR analysis of Bud13p ligand (residues 222–256) binding to Snu17p.** *a*, overlay of the $^{1}H,^{15}N$ HSQC spectra of the Bud13p ligand free and when bound to Snu17p. *b,* secondary ^{13}C chemical shifts $\Delta\delta(^{13}C^{\alpha}$-$^{13}C^{\beta})$ and $\{^{1}H\}$-^{15}N heteronuclear NOE data for the Bud13p ligand when bound to Snu17p. The last 5 residues of Bud13p have negative NOE values and are not shown. The short helix in the C-terminal region of the peptide is indicated on *top*.

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

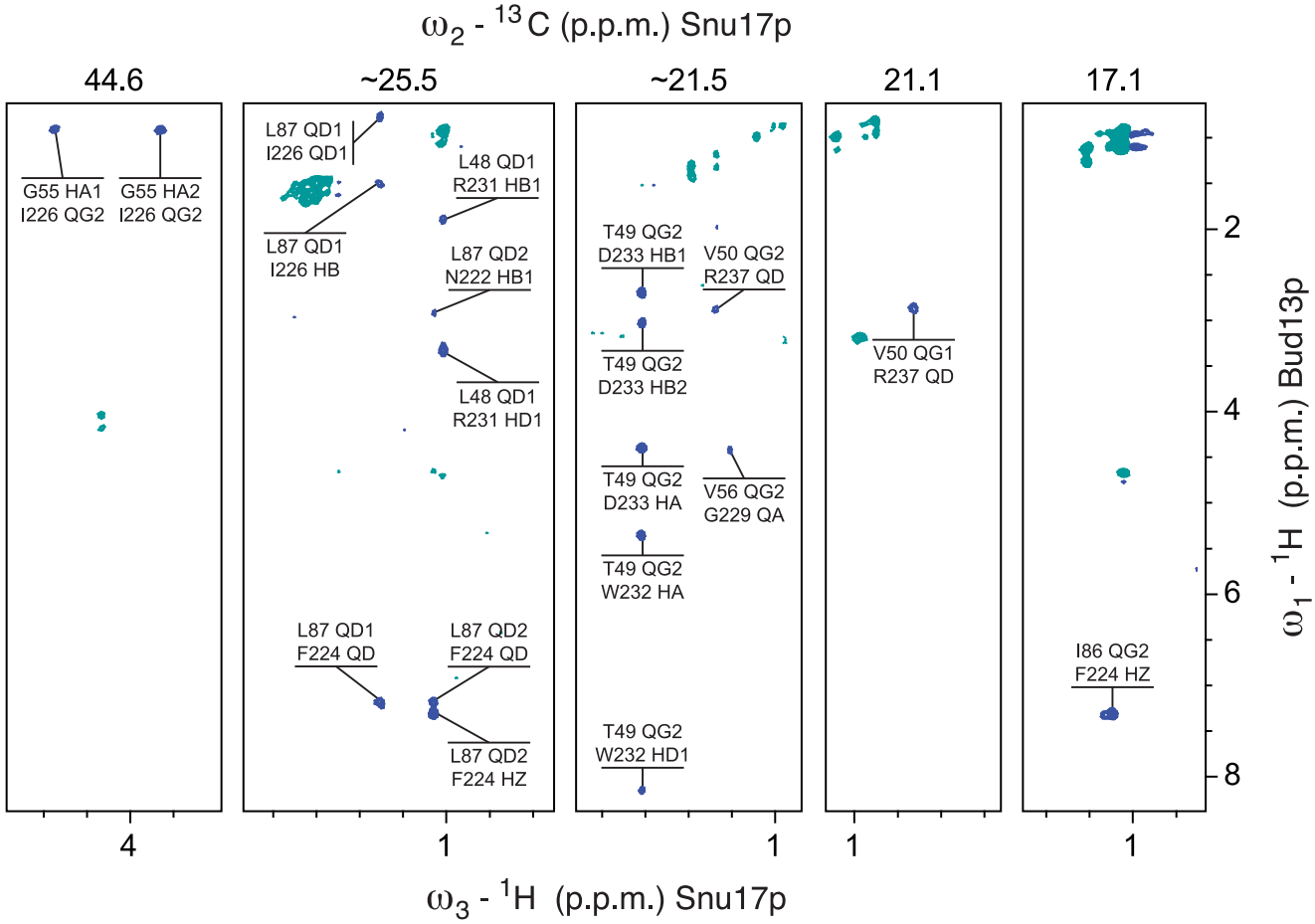


FIGURE 5. **^{1}H (ω_3),^{1}H (ω_1) planes from $^{13}C/^{15}N$-filtered, ^{13}C-edited NOESY-HSQC spectra in the binary complex composed of doubly labeled Snu17p and unlabeled Bud13p.** The ^{13}C chemical shifts (ppm) of the selected ω_2 planes are indicated on *top*. The experiment selects NOE cross-peaks between protons bound to ^{12}C or ^{14}N in Bud13p and protons bound to ^{13}C in Snu17p, providing exclusively intermolecular restraints. For each cross-peak the upper and lower atom name refers to Snu17p and Bud13p assignments, respectively.

NOESY-HSQC spectra (Fig. 5). The structure of the complex is well defined and based on 2607 experimental distance restraints (Fig. 6). A summary of the structural and restraint statistics is given in Table 1. The ensemble of the 20 lowest energy structures obtained after water refinement is shown in Fig. 6*a*.

In the complex Snu17p adopts an RRM-fold with two α-helices packing on one surface of an antiparallel β-sheet (Fig. 6*b*). The Snu17p structure superimposes very well with UHM family members (root mean square deviations 1.5–1.8 Å for 72 C_α atoms against the proteins listed in Fig. 1*a*). Noteworthy, although most UHM family members lack the conserved aromatic residues in RNP1 and RNP2 that mediate RNA binding in canonical RRMs, Snu17p has retained both RNPs. However, the structure shows that Tyr^{76} (RNP1) is not solvent exposed but rather stacks parallel with another tyrosine (Tyr^{109}) located at the C-terminal extension of the RRM-fold (Fig. 6*c*). The partial solvent protection of the RNP1 motif may imply potential regulation of RNA binding by Snu17p in the RES complex.

Bud13p binds to the helical surface of the RRM domain of Snu17p, consistent with chemical shift mapping data (Fig. 6*b*). Although Bud13p contains the invariant Trp residue of ULMs, the interaction is distinct from the canonical UHM-ULM recognition. Bud13p assumes an extended conformation and forms an arch around the Snu17p RRM above the two Snu17p helices. Two glycine residues flanking the invariant Trp that are not present in other ULMs change the backbone direction of the Bud13p on the surface of Snu17p. Eighteen amino acids of Bud13p interact with Snu17p forming a U-turn that crosses the helical interface of Snu17p (Fig. 6*b*). The invariant Trp is located in the middle of the turn and is recognized together with the flanking interacting residues of Bud13p on the surface of Snu17p (Fig. 6*b*). The C-terminal helix of the Bud13p ligand does not contribute to the binding interface and its orientation is not defined by experimental data. This suggests that it is flexible in solution, consistent with reduced heteronuclear $\{^{1}H\}$-^{15}N NOE values (Fig. 4*b*). The Snu17p-Bud13p binding surface area covers ~800 $Å^2$, and is thus significantly larger compared with any UHM-ULM interaction described (Fig. 6*d*).

The binding interface between Snu17p and Bud13p is mainly stabilized by hydrophobic interactions. The extensive bipartite hydrophobic interface includes the whole stretch of residues from Asn^{222} to Asn^{239} of Bud13p and the dorsal helical face of the Snu17p RRM domain. The apolar contacts are supported by electrostatic interactions involving mainly backbone amide groups of Bud13p (Fig. 6*d*). The entire backbone of the Bud13p stretch is involved in intermolecular contacts but certain amino acids engage their side chains as well, thus contributing more to

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

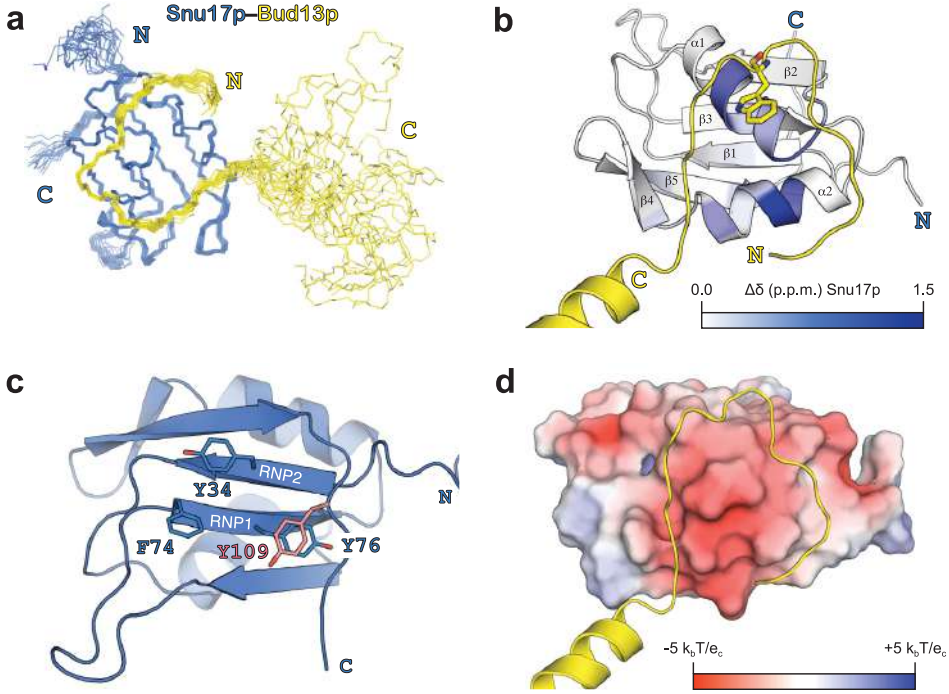


FIGURE 6. **Structure of the Snu17p-Bud13p complex.** *a*, overlay of the 20 lowest energy structures of Snu17p (*blue*) in complex with Bud13p (*yellow*). The N and C terminus of each protein is annotated. *b,* schematic representation of the Snu17p-Bud13p complex with amide CSP induced by Bud13p binding mapped onto the structure with a color gradient from *white* to *blue* with increasing CSP. The invariant Trp residue of Bud13p is shown as stick representation. *c*, the RNP signature motifs of Snu17p (Tyr^{34}, Phe^{74}, Tyr^{76}) are conserved but not entirely exposed to the solvent. *d,* electrostatic surface of Snu17p with negative and positive charges colored *red* and *blue*, respectively.

TABLE 1
NMR and refinement statistics for the Snu17p-Bud13p complex

NMR distance and dihedral restraints	
Distance restraints	
Total NOE	2607
Intra-residue	636
Inter-residue	1971
Sequential ($\|i-j\|=1$)	682
Medium range ($\|i-j\|<4$)	413
Long range ($\|i-j\|>5$)	747
Intermolecular	129
Total dihedral angle restraints	177
ϕ	87
ψ	90
Structure statistics	
Violations (mean ± S.D.)	
Distance restraints (Å)	0.011 ± 0.00
Dihedral angle restraints (°)	0.767 ± 0.07
Maximum dihedral angle violation (°)	5.79
Maximum distance restraint violation (Å)	0.34
Deviations from idealized geometry	
Bond lengths (Å)	0.011 ± 0.00
Bond angles (°)	1.206 ± 0.03
Impropers (°)	1.297 ± 0.06
Average pairwise root mean square deviation[a] (Å)	
Heavy	1.15 ± 0.17
Backbone	0.55 ± 0.12
Ramachandran plot statistics (%)	
Residues in most favored regions	93.4
Residues in additionally allowed regions	6.6

[a] Pairwise root mean square deviation was calculated among 20 refined structures for residues 30–109 of Snu17p and 222–239 of Bud13p.

the interaction. Three regions of higher significance can be discerned based on intermolecular NOE data and the relative contribution to the total interaction surface (Fig. 7*a*). The first spot is at the N terminus of Bud13p where it enters the interaction surface and passes above helix α2 of Snu17p. Phe^{224} and Ile^{226} of Bud13p exhibit mainly hydrophobic contacts with the helical interface lined by Snu17p Arg^{83}, Ser^{84}, Ile^{86}, and Leu^{87} (Fig. 7*b*). The second binding hot spot is at the U-turn of Bud13p. Bud13p residues from Pro^{228} to Val^{235} encircle the indole ring of the invariant Bud13p Trp^{232}. The curved conformation of the Bud13p peptide is supported by many intramolecular NOEs between Bud13p Trp^{232} and Pro^{228} on one side and Val^{235} on the other side. The presence of Gly^{229} following Pro^{228} and Gly^{234} preceding Val^{235} provide the conformational flexibility required for reversing the backbone course of Bud13p. In the U-turn configuration Pro^{228}, Arg^{231}, Trp^{232}, and Val^{235} of Bud13p form a continuous surface that interacts with Snu17p helix α1 (residues Leu^{48}, Thr^{49}, Ser^{52}, and Glu^{53}), as well as Snu17p Val^{56} in the linker following helix α1 (Fig. 7*b*). The third interface involves Snu17p Val^{50}, Phe^{95}, and Asn^{91}, which contact Bud13p Arg^{237}, Ser^{238}, and Asn^{239}, respectively, at the crevice between the two α-helices of the Snu17p RRM (Fig. 7*b*). This interface is further stabilized by hydrogen bonds between the Bud13p backbone and Snu17p side chains. At this point Bud13p ligand exits the binding interface with Snu17p and assumes a helical conformation.

DISCUSSION

The solution structure of Snu17p-Bud13p reveals a novel binding mode that has not been observed in other RRM-UHM peptide complexes. The structural findings are consistent with our initial mutagenesis analysis, which was guided by the recognition features described for the UHM family. We found that a double or triple mutant around the invariant Trp residue of Bud13p (R231D,W232A or R231D,W232A,D233N) disrupted Snu17p binding *in vitro*, and impacted on normal cell growth of transformants, whereas the single W232A substitution did not

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

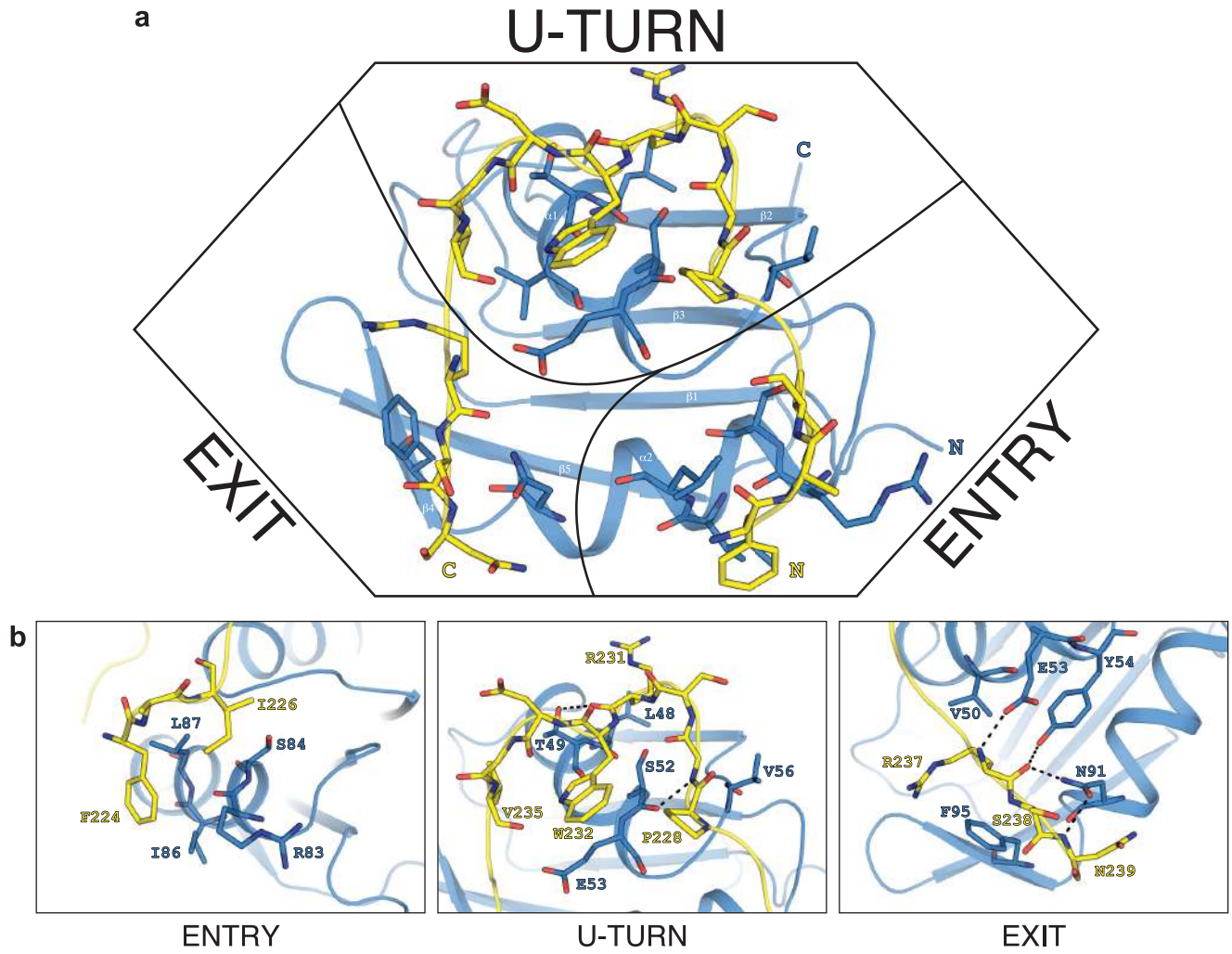


FIGURE 7. **Details of the intermolecular contacts in the Snu17p-Bud13p complex.** *a,* the Bud13p peptide forms an arch around the helical binding surface of Snu17p. *b,* close-up views of the interaction at the entry point (*left*), the U-turn (*middle*), and the exit point (*right*) of the Bud13p peptide shown in *a*. Interacting residues are annotated and colored *blue* for Snu17p and *yellow* for Bud13p. Intermolecular hydrogen bonds to the Bud13p backbone are indicated by *dashed lines*.

prevent complex formation (4), in contrast to canonical UHM-ULM interactions (18, 30). On the other hand, a triple mutant of Snu17p residues that correspond to the UHM Arg-*X*-Phe motif (F95A,K96D,I97A) had no effect in formation of the RES complex and restored a growth phenotype similar to wild type. These findings indicated an unexpected architecture of recognition that is distinct from the established UHM-ULM mode.

Several structures have revealed the main features that determine UHM-ULM interactions (16–19, 30, 31). UHM domains typically lack one or more aromatic residues in their RNP sequences, they have an acidic helix α1, and contain an Arg-*X*-Phe motif following the second helix α2. The Arg-*X*-Phe motif recognizes ULM sequences that contain an invariant Trp residue that is usually preceded by basic residues. The critical Trp is inserted into a hydrophobic cleft formed between the two helices of the UHM-fold with the ULM peptide oriented perpendicular to the two helices (Fig. 8*a*). In general, UHM domains do not interact with RNA as often the RNA-binding β-sheet surface on the opposite side of the RRM-fold is occluded by a C-terminal helix.

An interesting variation of the UHM-ULM theme is the interaction between the PTB protein and its corepressor Raver1 (32). The consensus signature motif of Raver1 ligands do not harbor the Trp residue but instead a pair of Leu residues. The PTB structures in complex with two peptide motifs from Raver1 (33) show that the Leu residues substitute for the Trp recognition seen in UHM-ULM binding and interact with a shallow groove formed at the same location on PTB RRM2 domain (Fig. 8*b*). Interestingly, it has been reported that the PTB RRM2 domain can bind simultaneously Raver1 and RNA, a distinctive feature that has not been described for other UHM domains.

The most notable exception of RRM binding to Trp containing ligands has been reported for the cellular mRNA export factor ALYREF (34). Two structures have been determined of the ALYREF RRM domain in complex with viral peptides, both different from the UHM-ULM mode (Fig. 8*c*). The ICP27 peptide binds ALYREF RRM in an extended conformation along the groove formed by the helices (34). The Trp residue is located at the upper edge of the first ALYREF helix and contrib-

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

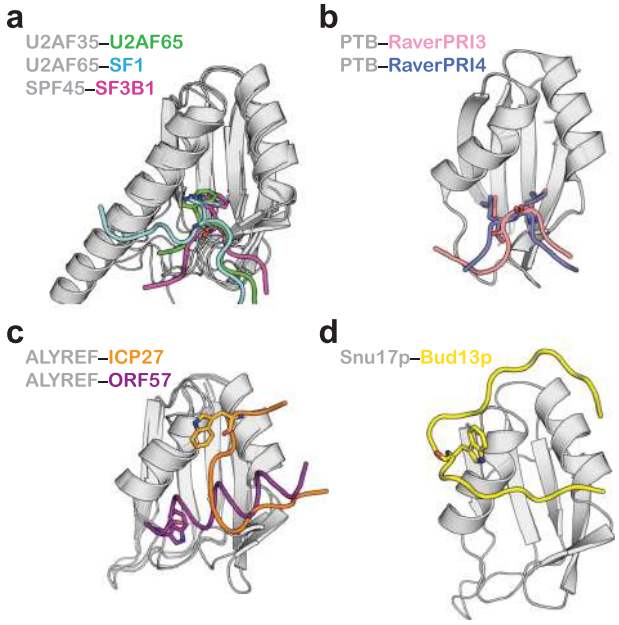


FIGURE 8. **Comparison of ligand binding modes of the UHM/RRM-fold.** In all panels the UHM/RRM domains are shown in the same orientation. *a,* three examples of the canonical UHM-ULM interaction. *b,* PTB binding to Raver1 ligands, which lack a Trp residue. *c,* two different binding modes of peptide partners with a Trp residue by ALYREF protein. *d,* the novel Bud13p peptide recognition by Snu17p.

utes to the total interaction surface. The ORF57 peptide on the other hand adopts a helical conformation running almost perpendicular to the helices of ALYREF (35). The Trp residue is positioned at one end of the helix and makes hydrophobic contacts with ALYREF along with other residues of the helical peptide. Mutational analysis indicates that in ORF57 the Trp residue is important for ALYREF recognition, whereas in ICP27 it is not as crucial as has been demonstrated for true UHM-ULM pairs.

The structure of Snu17p-Bud13p complex presented here displays yet another distinct binding topology that has not been described before, further adding to the diversity of RRM-ligand interactions. A large portion of the Bud13p ligand is engaged in contacts around the helical surface of Snu17p. The Trp residue is located in the middle of a U-turn configuration that collectively contacts the first helix of Snu17p (Fig. 8*d*). Preceding and following residues to the U-turn are perpendicular to the second helix of Snu17p and contribute to binding. The importance of these residues to the interaction, which are rather distant to the invariant Trp, has been confirmed by *in vitro* binding studies designed to identify the minimal interacting region on Bud13p. Notably, key residues in Bud13p involved in Snu17p binding are conserved (data not shown) in other orthologs including the human homolog of Bud13p, suggesting that the novel mode of RRM-peptide interaction is conserved.

The yeast RES complex facilitates the splicing of some introns and participates in the retention of unspliced pre-mRNAs in the nucleus. The absence of a functional RES complex in yeast results in slow growth, particularly at high temperature, due to reduced splicing (3). Our data show that preventing the association of Bud13p with Snu17p impairs cell growth without affecting protein expression levels. This demonstrates that individual proteins when simultaneously present are unable to perform the RES activity and thus that assembly of the RES complex is essential for function. It is conceivable that to perform its task the RES complex recognizes specific RNA sequences through its Snu17p subunit. The Snu17p-Bud13p structure shows that Bud13p binding does not interfere with the RNP sequences of Snu17p. Snu17p could thus be involved in RNA recognition as shown for canonical RRM domains (13, 14). However, in Snu17p the first RNP motif is partially occluded from the solvent by the C-terminal tail of the RRM-fold. We envisage that Pml1p may regulate RNA binding of Snu17p in the functional ternary RES complex. The binding of Pml1p, which requires the Snu17p C-terminal portion in addition to the RRM-fold, could induce structural rearrangements that might completely expose the RNP motifs of Snu17p. Future studies with the ternary RES complex and RNA should shed light to this hypothesis and increase our understanding for RES function in RNA splicing and nuclear retention of pre-messenger transcripts.

Acknowledgments—We thank members of the Sattler and Séraphin groups for discussion and advice. We acknowledge NMR measurement time at the Bavarian NMR Centre, Garching, Germany.

REFERENCES

1. Jurica, M. S., and Moore, M. J. (2003) Pre-mRNA splicing: awash in a sea of proteins. *Mol. Cell* **12,** 5–14
2. Wahl, M. C., Will, C. L., and Lührmann, R. (2009) The spliceosome: design principles of a dynamic RNP machine. *Cell* **136,** 701–718
3. Dziembowski, A., Ventura, A. P., Rutz, B., Caspary, F., Faux, C., Halgand, F., Laprévote, O., and Séraphin, B. (2004) Proteomic analysis identifies a new complex required for nuclear pre-mRNA retention and splicing. *EMBO J.* **23,** 4847–4856
4. Brooks, M. A., Dziembowski, A., Quevillon-Cheruel, S., Henriot, V., Faux, C., van Tilbeurgh, H., and Séraphin, B. (2009) Structure of the yeast Pml1 splicing factor and its integration into the RES complex. *Nucleic Acids Res.* **37,** 129–143
5. Trowitzsch, S., Weber, G., Lührmann, R., and Wahl, M. C. (2009) Crystal structure of the Pml1p subunit of the yeast precursor mRNA retention and splicing complex. *J. Mol. Biol.* **385,** 531–541
6. Trowitzsch, S., Weber, G., Lührmann, R., and Wahl, M. C. (2008) An unusual RNA recognition motif acts as a scaffold for multiple proteins in the pre-mRNA retention and splicing complex. *J. Biol. Chem.* **283,** 32317–32327
7. Deckert, J., Hartmuth, K., Boehringer, D., Behzadnia, N., Will, C. L., Kastner, B., Stark, H., Urlaub, H., and Lührmann, R. (2006) Protein composition and electron microscopy structure of affinity-purified human spliceosomal B complexes isolated under physiological conditions. *Mol. Cell. Biol.* **26,** 5528–5543
8. Bessonov, S., Anokhina, M., Will, C. L., Urlaub, H., and Lührmann, R. (2008) Isolation of an active step I spliceosome and composition of its RNP core. *Nature* **452,** 846–850
9. Dreumont, N., and Séraphin, B. (2013) Rapid screening of yeast mutants with reporters identifies new splicing phenotypes. *FEBS J.* **280,** 2712–2726
10. Scherrer, F. W., Jr., and Spingola, M. (2006) A subset of Mer1p-dependent introns requires Bud13p for splicing activation and nuclear retention. *RNA* **12,** 1361–1372
11. Zhou, Y., Chen, C., and Johansson, M. J. (2013) The pre-mRNA retention and splicing complex controls tRNA maturation by promoting TAN1 expression. *Nucleic Acids Res.* **41,** 5669–5678
12. Hausmann, S., Zheng, S., Costanzo, M., Brost, R. L., Garcin, D., Boone, C., Shuman, S., and Schwer, B. (2008) Genetic and biochemical analysis of yeast and human cap trimethylguanosine synthase: functional overlap of

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

2,2,7-trimethylguanosine caps, small nuclear ribonucleoprotein components, pre-mRNA splicing factors, and RNA decay pathways. *J. Biol. Chem.* **283,** 31706–31718
13. Daubner, G. M., Cléry, A., and Allain, F. H. (2013) RRM-RNA recognition: NMR or crystallography and new findings. *Curr. Opin. Struct. Biol.* **23,** 100–108
14. Maris, C., Dominguez, C., and Allain, F. H. (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J.* **272,** 2118–2131
15. Kielkopf, C. L., Lücke, S., and Green, M. R. (2004) U2AF homology motifs: protein recognition in the RRM world. *Genes Dev.* **18,** 1513–1526
16. Corsini, L., Bonnal, S., Bonna, S., Basquin, J., Hothorn, M., Scheffzek, K., Valcárcel, J., and Sattler, M. (2007) U2AF-homology motif interactions are required for alternative splicing regulation by SPF45. *Nat. Struct. Mol. Biol.* **14,** 620–629
17. Corsini, L., Hothorn, M., Stier, G., Rybin, V., Scheffzek, K., Gibson, T. J., and Sattler, M. (2009) Dimerization and protein binding specificity of the U2AF homology motif of the splicing factor Puf60. *J. Biol. Chem.* **284,** 630–639
18. Selenko, P., Gregorovic, G., Sprangers, R., Stier, G., Rhani, Z., Krämer, A., and Sattler, M. (2003) Structural basis for the molecular recognition between human splicing factors U2AF65 and SF1/mBBP. *Mol. Cell* **11,** 965–976
19. Loerch, S., Maucuer, A., Manceau, V., Green, M. R., and Kielkopf, C. L. (2014) Cancer-relevant splicing factor CAPERα engages the essential splicing factor SF3b155 in a specific ternary complex. *J. Biol. Chem.* **289,** 17325–17337
20. Collinet, B., Friberg, A., Brooks, M. A., van den Elzen, T., Henriot, V., Dziembowski, A., Graille, M., Durand, D., Leulliot, N., Saint André, C., Lazar, N., Sattler, M., Séraphin, B., and van Tilbeurgh, H. (2011) Strategies for the structural analysis of multi-protein complexes: lessons from the 3D-Repertoire project. *J. Struct. Biol.* **175,** 147–158
21. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6,** 277–293
22. Sattler, M., Schleucher, J. R., and Griesinger, C. (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nucl. Magn. Reson. Spectrosc.* **34,** 93–158
23. Breeze, A. L. (2000) Isotope-filtered NMR methods for the study of biomolecular structure and interactions. *Prog. Nucl. Magn. Reson. Spectrosc.* **36,** 323–372
24. Neri, D., Szyperski, T., Otting, G., Senn, H., and Wüthrich, K. (1989) Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 repressor by biosynthetically directed fractional ^{13}C labeling. *Biochemistry* **28,** 7510–7516
25. Güntert, P. (2009) Automated structure determination from NMR spectra. *Eur. Biophys. J.* **38,** 129–143
26. Shen, Y., Delaglio, F., Cornilescu, G., and Bax, A. (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR* **44,** 213–223
27. Linge, J. P., Williams, M. A., Spronk, C. A., Bonvin, A. M., and Nilges, M. (2003) Refinement of protein structures in explicit solvent. *Proteins* **50,** 496–506
28. Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54,** 905–921
29. Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Séraphin, B. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17,** 1030–1032
30. Kielkopf, C. L., Rodionova, N. A., Green, M. R., and Burley, S. K. (2001) A novel peptide recognition mode revealed by the X-ray structure of a core U2AF35/U2AF65 heterodimer. *Cell* **106,** 595–605
31. Elantak, L., Wagner, S., Herrmannová, A., Karásková, M., Rutkai, E., Lukavsky, P. J., and Valásek, L. (2010) The indispensable N-terminal half of eIF3j/HCR1 cooperates with its structurally conserved binding partner eIF3b/PRT1-RRM and with eIF1A in stringent AUG selection. *J. Mol. Biol.* **396,** 1097–1116
32. Rideau, A. P., Gooding, C., Simpson, P. J., Monie, T. P., Lorenz, M., Hüttelmaier, S., Singer, R. H., Matthews, S., Curry, S., and Smith, C. W. (2006) A peptide motif in Raver1 mediates splicing repression by interaction with the PTB RRM2 domain. *Nat. Struct. Mol. Biol.* **13,** 839–848
33. Joshi, A., Coelho, M. B., Kotik-Kogan, O., Simpson, P. J., Matthews, S. J., Smith, C. W., and Curry, S. (2011) Crystallographic analysis of polypyrimidine tract-binding protein-Raver1 interactions involved in regulation of alternative splicing. *Structure* **19,** 1816–1825
34. Tunnicliffe, R. B., Hautbergue, G. M., Kalra, P., Jackson, B. R., Whitehouse, A., Wilson, S. A., and Golovanov, A. P. (2011) Structural basis for the recognition of cellular mRNA export factor REF by herpes viral proteins HSV-1 ICP27 and HVS ORF57. *PLoS Pathog.* **7,** e1001244
35. Tunnicliffe, R. B., Hautbergue, G. M., Wilson, S. A., Kalra, P., and Golovanov, A. P. (2014) Competitive and cooperative interactions mediate RNA transfer from herpesvirus saimiri ORF57 to the mammalian export adaptor ALYREF. *PLoS Pathog.* **10,** e1003907

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

5208–5220 *Nucleic Acids Research, 2015, Vol. 43, No. 10* *Published online 27 April 2015*
doi: 10.1093/nar/gkv365

A novel RNA binding surface of the TAM domain of TIP5/BAZ2A mediates epigenetic regulation of rRNA genes

Irina Anosova[1,2], Svitlana Melnik[3], Konstantinos Tripsianes[1,2], Fatiha Kateb[1,2], Ingrid Grummt[3,*] and Michael Sattler[1,2,*]

[1]Institute of Structural Biology, Helmholtz Zentrum München, Neuherberg D-85764, Germany, [2]Biomolecular NMR and Center for Integrated Protein Science Munich (CIPSM), Department Chemie, Technische Universität München, Garching D-85747, Germany and [3]Division of Molecular Biology of the Cell II, DKFZ-ZMBH Alliance, German Cancer Research Center, Heidelberg D-69120, Germany

Received November 19, 2014; Revised March 19, 2015; Accepted April 3, 2015

ABSTRACT

The chromatin remodeling complex NoRC, comprising the subunits SNF2h and TIP5/BAZ2A, mediates heterochromatin formation at major clusters of repetitive elements, including rRNA genes, centromeres and telomeres. Association with chromatin requires the interaction of the TAM (TIP5/ARBP/MBD) domain of TIP5 with noncoding RNA, which targets NoRC to specific genomic loci. Here, we show that the NMR structure of the TAM domain of TIP5 resembles the fold of the MBD domain, found in methyl-CpG binding proteins. However, the TAM domain exhibits an extended MBD fold with unique C-terminal extensions that constitute a novel surface for RNA binding. Mutation of critical amino acids within this surface abolishes RNA binding *in vitro* and *in vivo*. Our results explain the distinct binding specificities of TAM and MBD domains to RNA and methylated DNA, respectively, and reveal structural features for the interaction of NoRC with non-coding RNA.

INTRODUCTION

Despite the fact that mammalian cells contain several hundreds of rRNA genes (rDNA) and rRNA genes are the most actively transcribed genes in eukaryotes, a large fraction of rDNA is silenced by heterochromatin formation to prevent aberrant homologous recombination, thus safeguarding rDNA stability and nucleolar integrity (1). The transcriptionally silent state of rDNA is established by the chromatin remodeling complex NoRC, which comprises the DNA-dependent adenosine triphosphatase SNF2h and a large subunit, termed TIP5 (also known as BAZ2A) (2) (Figure 1a). The C-terminal part of TIP5 harbors a tandem PHD finger/bromodomain, a cooperative unit that interacts with chromatin modifying enzymes, which set specific heterochromatic histone marks and trigger *de novo* DNA methylation (3,4). Targeting NoRC to rDNA leads to repositioning of the promoter-bound nucleosome, changes in histone modifications, increased DNA methylation and silencing of rRNA genes (5–8). Consistent with epigenetic regulation representing an intimate and balanced interplay of both RNA and chromatin fields, NoRC function requires the association of TIP5 with specific 150–250 nt RNA, named pRNA ('promoter-associated RNA') as its sequence overlaps the rDNA promoter (6,9). pRNA folds into a phylogenetically conserved hairpin structure that is recognized by TIP5 and this specific secondary structure of the pRNA is required for guiding NoRC to nucleoli (10). Mutations that disrupt this specific stem-loop structure impair binding of TIP5 to pRNA and abolish the nucleolar localization of NoRC,whereas compensatory mutations that restore the hairpin structure re-establish binding and targeting NoRC to nucleoli. The interaction with pRNA is mediated by the TAM (TIP5/ARBP/MBD) domain in the N-terminal region of TIP5, and is indispensable for nucleolar localization of NoRC and heterochromatin formation at rDNA. RNase footprinting and protease sensitivity experiments suggest that TIP5 associates with pRNA in an induced fit mechanism, resulting in structural changes that may facilitate the interaction with co-repressors to promote heterochromatin formation and rDNA silencing (6). A recent report demonstrated that the poly (ADP-ribose)-polymerase-1 (PARP1) associates with TIP5, thus acting as a co-repressor that mediates the inheritance of silent histone marks through cell cycle progression (11). Significantly, NoRC function is not restricted to rDNA silencing, but it has been shown to

*To whom correspondence should be addressed. Tel: +49 89 289 13867; Fax: +49 89 289 13869; Email: sattler@helmholtz-muenchen.de
Correspondence may also be addressed to Ingrid Grummt. Tel: +49 6221 423423; Fax:+49 6221 422995; Email: I.Grummt@dkfz.de
Present address: Konstantinos Tripsianes, Central European Institute of Technology (CEITEC), Masaryk University, Kamenice 5, 62500 Brno, Czech Republic.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

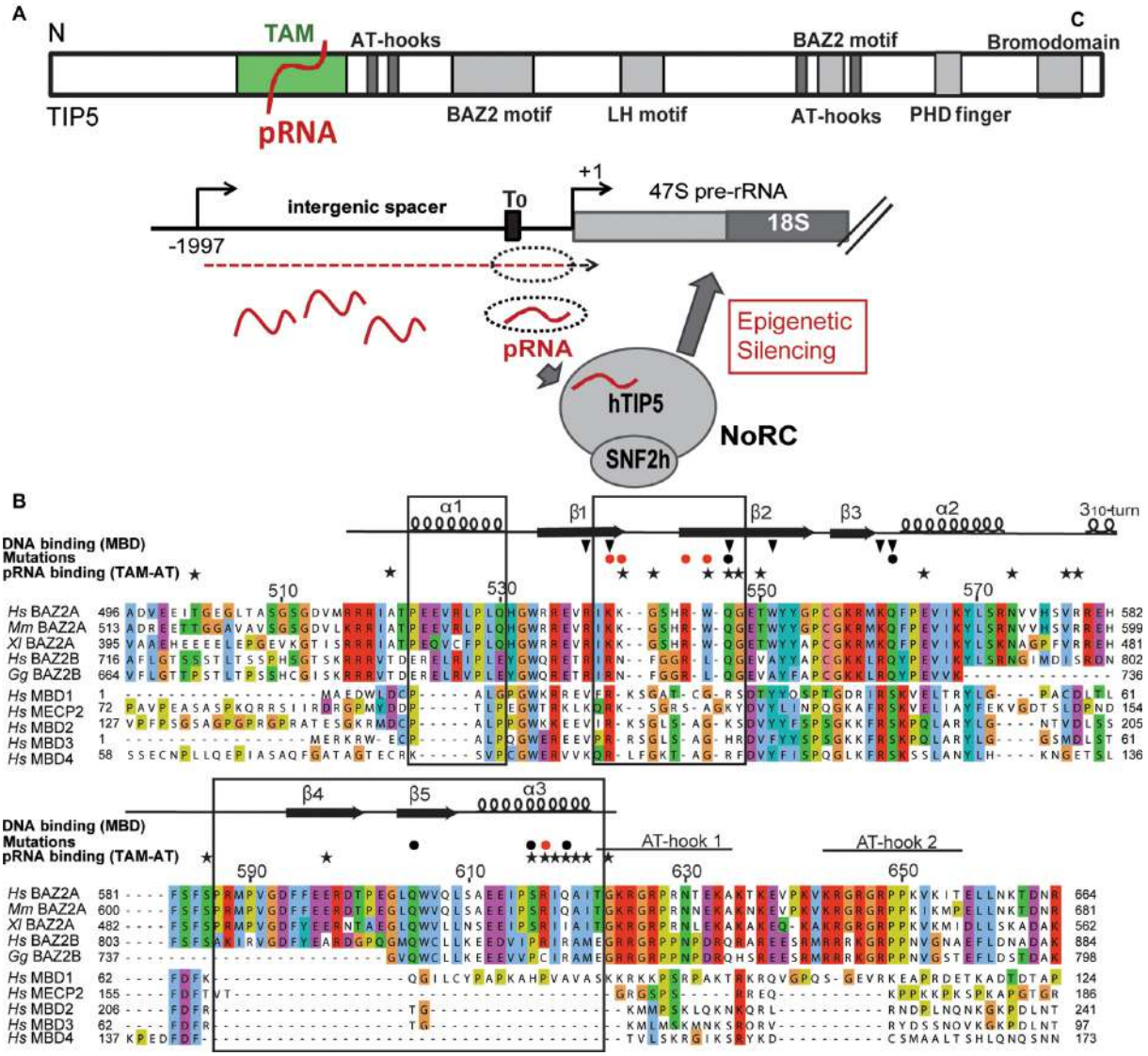


Figure 1. Domain organization of TIP5 and its TAM domain. (**A**) Top: domain organization of human TIP5. Bottom: the role of pRNA in NoRC-mediated silencing of rRNA genes (1). Intergenic transcripts (red dashed line) originating from the rDNA intergenic spacer are processed into 150–250 nt transcripts, termed pRNA (red waved lines). The interaction of pRNA with TIP5 guides NoRC to the rDNA promoter, an essential step in NoRC-dependent silencing of rRNA genes. (**B**) Multiple sequence alignment of the TAM domain of hTIP5 (BAZ2A) with phylogenetically related BAZ2A homologs (14) and canonical human MBDs. Sequences were aligned using MUSCLE algorithm (47), color coding according to CLUSTAL-X (48). Secondary structure observed in human TAM (Figure 3) is indicated above the sequence. Structural features of the TAM domain that are not present in MBD folds are indicated by black boxes. Black stars mark residues that are strongly affected in NMR titrations of TIP5/TAM-AT with $pRNA^{mini}$. Red and black filled circles indicate residues that were mutated to probe the TAM/pRNA interaction interface showing strong or no effect on RNA binding, respectively. Black triangles indicate key residues of the MBD1 that are crucial for its interaction with methylated CpG (52). Many of these residues are not conserved in TAM domains, i.e. TIP5 TAM R538, Q547, Q562. Residue numbering refers to human TIP5.

also establish a repressive heterochromatic structure at centromeres and telomeres, thus preserving the structural integrity of these repetitive genomic loci (12). Paradoxically, TIP5 may serve as a useful marker for the metastatic potential in prostate cancer. By interacting with EZH2, TIP5 hypermethylates and silences genes that are repressed in metastasis (13).

The TAM domain in the N-terminal region of TIP5 exerts sequence homology to methyl-CpG-binding domains (MBDs), which are found in proteins that bind to methylated DNA and mediate methylation-dependent gene silencing (14). TIP5 also harbors two AT-hooks C-terminally of the TAM domain, which are known to bind DNA and regulate transcription by modifying the architecture of DNA (15). Canonical MBDs bind to methylated cytosine in the context of CpG motifs in DNA but do not directly bind RNA (16). The closely related TAM domain of TIP5, on the other hand, directly contacts pRNA and this interaction is essential for targeting NoRC to rDNA and NoRC-dependent silencing of rRNA genes (10).

Here, we report the three-dimensional (3D) solution structure of the TAM domain of human TIP5 (hTIP5).

We show that the TAM domain represents a variant of the MBD fold with unique C-terminal extensions that are required for the interaction of TIP5 with pRNA. Using biochemical studies and NMR-spectroscopy we show that point mutations in this surface impair RNA binding *in vitro* and *in vivo*. Our results provide structural details for interactions between TIP5 and RNA that are essential for recruitment of the chromatin remodeling complex NoRC to specific genomic loci.

MATERIALS AND METHODS

Plasmid constructs, cells, general procedures

HEK293T cells were grown in DMEM (Invitrogen) supplied with 10% fetal calf serum (FCS; Biosera), 11 mg/ml sodium pyruvate and 11 mg/ml penicillin and streptomycin (PAA Laboratories GmbH). DNA encoding human TIP5/TAM-AT (residues 496–664) and human TIP5/TAM (residues 516–623) were subcloned from pJC40-hTIP5(N332–723) using standard cloning protocols into modified pETM-11 vector (EMBL) containing a His_6-tag for purification, a GB1-tag at the N-terminus for protein solubility and a tobacco etch virus cleavage site for removal of the tags. TIP5 variants were obtained by QuickChange site-directed mutagenesis (Agilent Technologies). pJC40-hTIP5(N332–723) and pJC40-mTIP5(N1–598) used for purification of human or murine TIP5 for *in vitro* studies were described before (6). HA-FLAG-hTIP5 containing full-length human TIP5 or TIP5 mutants used for *in vivo* experiments was described before (2). pTOPO2.1-mrDNA (-204/-1 sense) used for the synthesis of mouse pRNA was described before (6).

Protein expression and purification

Isotopically labeled proteins for NMR studies were produced by overexpression in *Escherichia coli* BL21 (DE3) cells in minimal medium (M9) enriched with ^{15}N, ^{13}C and/or ^{2}H, supplemented with 50μg/ml kanamycin. $^{15}N/^{13}C$- and ^{15}N-labeled proteins were prepared using M9 medium with [U-^{13}C]-D-glucose (2g/l) and/or $^{15}NH_4Cl$ (1g/l) as sole carbon and nitrogen sources. For 50% random fractional deuteration the M9 medium was based on 90% $^{2}H_2O$. Media were inoculated overnight and grown to $OD_{600} = 0.8$ at 37°C. Expression was induced with 0.2 mM IPTG and performed at 18°C for at least 16 h.

Escherichia coli transformed with cDNA encoding wild-type or mutant TIP5 were resuspended in lysis buffer containing 50 mM HEPES [pH 8.0], 1 M NaCl, 10 mM imidazole, 2 mM β-mercaptoethanol, 1 mM Pefabloc SC (Serva), 1 μg/ml lysozyme and 7μg/ml DNase I, 0.2% w/v IgepalCA-630. After sonication and high-speed centrifugation, His-tagged proteins were bound to Ni-NTA Agarose (Quiagen). Insoluble TIP5/TAM was purified from the pellet by resuspending in low salt lysis buffer with 6 M guanidinium hydrochloride and refolding on a Ni^{2+} column by washing with decreasing concentrations of guanidinium hydrochloride (2–0 M). After washing with low salt and high salt buffer containing 1.8 M NaCl, bound proteins were eluted with 300 mM imidazole, digested with tobacco etch virus protease (1 μg/mg protein at 4°C overnight). The tags were removed with a second Ni^{2+}column and recombinant proteins were further purified by size-exclusion chromatography (HiLoad 26/60 Superdex 75, GE Healthcare) in low salt lysis buffer. For NMR measurements proteins were kept in 20 mM sodium phosphate [pH 7.0], 50 mM NaCl, 1 mM Dithiothreitol (DTT), 10% v/v $^{2}H_2O$.

RNA transcription and complex formation

$pRNA^{mini}$ was transcribed *in vitro* by T7 RNA polymerase from a double stranded synthetic DNA template (17,18) using 12 mM $MgCl_2$, 4 mM adenosine triphosphate (ATP) and cytidine triphosphate (CTP), 4.8 mM guanosin-5'-triphosphate (GTP) and uridine-5'-triphosphate (UTP) and 0.08mg/ml PEG8000 (Promega). For reduction of $n + 1$ products the bottom strand was methoxylated at two 5′-terminal nucleotides (19). Transcripts were precipitated with ethanol, purified by preparative 20% (19:1) PAGE and electroeluted at 4°C. Residual acrylamide was removed by dialysis in 10 mM sodium phosphate [pH 6.0], 0.02 mM ethylenediaminetetraacetic acid (EDTA), 0.02 mM NaN_3 at 1-, 0.5- and 0 M NaCl. Formation of the stem-loop in solution was induced by denaturing $pRNA^{mini}$ in NMR buffer for 4 min at 95°C and snap-cooling on ice for 15 min and controlled by NMR.

Complex formation was performed at low concentrations on ice. The required amount of $pRNA^{mini}$ in 1.5 ml of cold NMR buffer was slowly added to 33 nmol protein in 2.5 ml of cold NMR buffer. The solution was gently mixed, left for 20 min on ice to allow complex formation and concentrated in an Amicon® Ultra centrifugal filter unit (Merck Millipore) to yield a complex of 0.15 mM TIP5/TAM-AT with molar equivalents of $pRNA^{mini}$. NMR data were recorded on the complex in a Shigemi tube directly thereafter.

NMR spectroscopy

NMR measurements were carried out at 298K (TIP5/TAM-AT) and 293K (TIP5/TAM) on Bruker 900-, 750- or 600 MHz spectrometers equipped with room temperature (750 MHz) or cryoprobes (900, 600 MHz) and pulsed field gradients. Spectra were processed with NMRPipe (20) and analyzed using Sparky 3.110 software (T. D. Goddard, D. G. Kneller, SPARKY 3, University of California, San Francisco). Backbone resonances were derived using a standard set of 3D NMR experiments for backbone assignment (21) (HNCA, HNCACB, HN(CO)CACB and CBCA(CO)NH) on $^{15}N/^{13}C$ uniformly labeled and in case of TIP5/TAM-AT 50% randomly deuterated protein samples. Secondary chemical shifts were calculated and used for secondary structure prediction as described previously (22,23). ^{1}H and ^{13}C side chain resonances were assigned based on 3D HCCH-TOCSY experiments with ^{13}C and ^{1}H evolution and correlated to amide group by CC(CO)NH-TOCSY. $^{13}C\beta$ and $^{1}H\delta/\epsilon$ of aromatic residues were correlated via scalar couplings (24). The tautomeric state of histidines was determined from long range $^{1}H,^{15}N$-HMQC (25). Stereospecific assignments of pro-S and pro-R methyl ^{1}H resonances of valine and leucine were based on a 10% ^{13}C labeled NMR sample as described (26,27). Backbone- and side chain residues

were assigned manually supported by the MARS software (28). NOE-assignments were performed using CYANA 3.0 (29) and checked manually. Torsion angle restraints were derived from TALOS+ (30).

$^{1}H^{N}-^{15}N$ and $^{13}CO-^{15}N$ RDC were recorded in a two-dimensional doublet-separated, sensitivity enhanced HSQC experiment (31). The samples were partially aligned using 7% PAA compressed gel (32) or 4% w/v 3:1 DMPC/DHPC bicelles (33) (Avanti Polar Lipids).

Relaxation measurements were performed at 900 MHz proton Larmor frequency on a 0.2 mM ^{15}N-labeled sample of TIP5/TAM-AT at 298K and at 750 MHz proton Larmor frequency on a 0.6 mM $^{15}N/^{13}C$ labeled sample of TIP5/TAM at 293K. R_1, $R_{1\rho}$ and R_2 were measured in an interleaved fashion as described (34). The R_2 relaxation rate was derived from $R_{1\rho}$ (35) for TIP5/TAM-AT and measured directly for the TIP5/TAM domain. NMR spectra were processed and analyzed with NMRPipe (20) using NMRDraw for visualization. Peak intensities were extracted at different delays from peak volume for non-overlapping signals. The relaxation rates were fitted to a 2-parameter exponential decay from the intensities, with errors from duplicate experiments, calculated from both the standard deviation and the fit itself. The $\{^{1}H\}$-^{15}N heteronuclear NOE was extracted as the ratio of peak intensities recorded with and without saturation. The error was estimated from intensity of experimental noise with error propagation. Analysis of the relaxation data was performed as described previously (36). The tumbling correlation time (τ_c) was estimated from R_2/R_1 (37) and compared to the expected value, calculated from the number of residues in a monomer (38). Relaxation data were plotted using Xmgrace.

Proton/deuterium exchange data were measured at 600 MHz Larmor frequency on a 0.2 mM ^{15}N-labeled sample at 293K. The protein sample was split in half, one being used for setting up experimental parameters on the spectrometer, the other being lyophilized and redissolved in 100% $^{2}H_2O$ immediately before measurement. The exchange was detected by the reduction/disappearance of the respective $^{1}H^{N}$-^{15}N signals in a series of SOFAST HMQCs (39) recorded at time intervals ranging from 8–120 min. The experiment was recorded with 8 scans, taking ca. 5–6 min per measurement.

NMR chemical shift perturbation experiments were measured at 600 MHz Larmor frequency on a complex of 0.15 mM ^{15}N-labeled, 50% deuterated TIP5/TAM-AT with different molar ratios of unlabeled $pRNA^{mini}$ at 298K. Titrations of TIP5/TAM with additional RNA and DNA constructs were measured on 0.1–0.15 mM ^{15}N-labeled and ^{15}N- and 10% ^{13}C-labeled samples at 293 K and Larmor frequency of 600 MHz and 500 MHz, respectively. Nucleic acids were added at increasing molar ratios to at least equimolar amounts. Chemical shift changes of $^{1}H^{N}$-^{15}N signals of the protein were monitored in a series of ^{1}H, ^{15}N HSQC experiments. The peak positions were determined using the SPARKY 3.110 automatic picking algorithm. Chemical shift changes are calculated as:

$$\Delta\delta_{ppm} = \sqrt{(\delta H_x - \delta H_0)^2 + \left(\frac{\delta N_x - \delta N_0}{5}\right)^2},$$

where δH_x and δN_x are the respective chemical shift values of a resonance in the complex; δH_0 and δN_0 are the respective chemical shift values of the same resonance in the free protein. The factor 1/5 was used to account for the different ppm spectral range of proton and nitrogen spins. Resonances with chemical shift changes or intensity drops larger than one standard deviation above the average were considered significant for binding.

Structure determination

The solution structure of the TIP5/TAM domain was defined based on 2732 distance restraints derived from CYANA 3.0 calculations, torsion angle restrains and two sets of RDCs measured in different alignment media. Two hundred structures were calculated and water-refined by ARIA1.2/CNS (40,41). The 20 lowest-energy structures were used for quality and structure validation by the iCING web interface (42), PROCHECK (43) and WHATCHECK (44). Backbone torsion angles for 92.1, 7.8 and 0.1% of all residues are found in the core, allowed and generally allowed regions of the Ramachandran plot, respectively, none in disallowed regions. The RDC R-factor was calculated as described previously (45). Electrostatic potential maps were calculated online using PyMOL (The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger, LLC). Structural representations were generated with PyMOL.

Electrophoretic mobility shift assays (EMSA)

Electrophoretic mobility shift assays (EMSA) experiments were performed as described previously (6). Radiolabeled MCS-RNA was synthesized by T7 RNA polymerase using pBluescript-KS linearized with EcoRI and transcribed with T7 RNA polymerase in the presence of 50 μCi [^{32}P]-UTP (3,000 Ci/mmol; Perkin Elmer). Radiolabeled RNA was incubated for 30 min on ice with purified recombinant protein (10–50 nM) in 12 mM Tris–HCl, pH 8.0, 100 mM KCl, 5mM $MgCl_2$, 0.1 mM EDTA, 0.5 mM DTT, 3% glycerol. In competition experiments (Supplementary Figure S2), unlabeled competitor RNA was added after 5 min and reactions were incubated for additional 30 min. RNA–protein complexes were resolved in 6% native polyacrylamide gels in 0.5× Tris/Borate/EDTA buffer (TBE). Gels were vacuum-dried, exposed to a screen and visualized by PhosphorImaging. The amount of bound RNA in band shifts and unbound RNA probes was compared by quantification of images using Image Gauge 4.0 software.

Filter binding assays (FBA)

Filter binding assays (FBA) were performed essentially as described (46).

For synthesis of [^{32}P]-labeled pRNA, pTOPO2.1-mrDNA(-204/-1) was linearized with EcoRI and transcribed with T7 RNA polymerase in the presence of 50

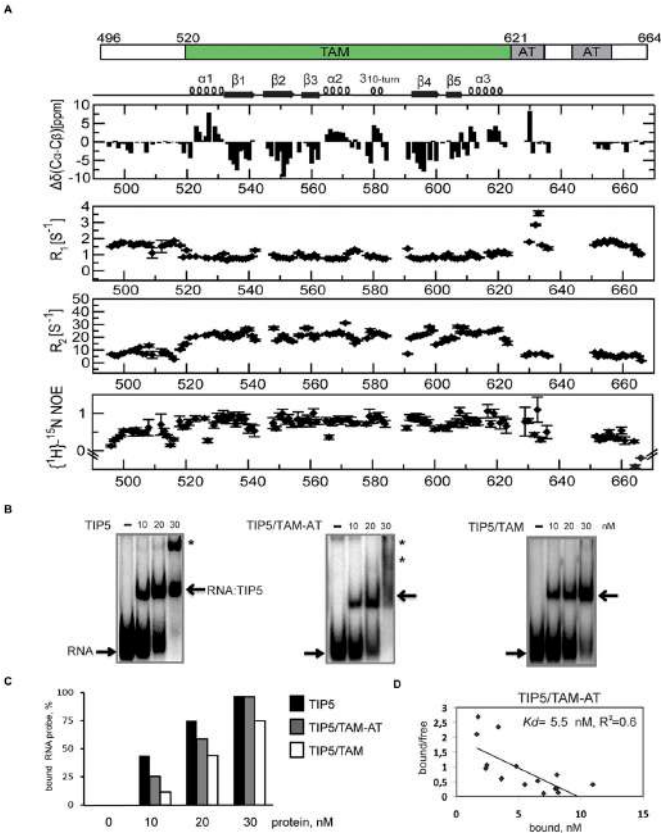


Figure 2. NMR and biochemical analysis of RNA binding by the TIP5/TAM domain. (**A**) NMR analysis of the TIP5/TAM-AT domain. ^{13}C secondary chemical shifts $\Delta\delta(^{13}C_\alpha) - \Delta\delta(^{13}C_\beta)$, ^{15}N R_1, R_2 relaxation rates and $\{^1H\}$-^{15}N heteronuclear NOE values are plotted versus TIP5 residue numbers. Secondary structure elements of the TAM domain are indicated on top. (**B**) The TAM-AT domain represents the minimal RNA binding region of TIP5. Electrophoretic mobility shift assays (EMSA) showing binding of TIP5, TIP5/TAM and TIP5/TAM-AT to RNA. 0, 10, 20 and 30 nM of the respective proteins were incubated with 1 nM of [^{32}P]-MCS-RNA probe. Positions of free RNA probe and RNA:TIP5 complexes are marked with arrows. Super-shifted binding products are marked with asterisks. (**C**) TIP5/TAM-AT has almost the same RNA binding efficiency as TIP5. Amounts of bound RNA in the EMSAs shown in panel (B) were quantitated using the ImageGauge software and are presented as percentage of input probe. (**D**) Scatchard plot analysis of *in vitro* binding of TIP5/TAM-AT to pRNA. Binding kinetics to pRNA were monitored by filter binding assay. The dissociation constant (K_D) was calculated using Scatchard plot analysis from the ratio of bound to bound/free probes. The resulting fit ($R^2 = 0.6$) yields an RNA binding affinity constant for the TIP5/TAM-AT/pRNA interaction of $K_D = 5.5$ nM.

μCi [^{32}P]UTP (3000 Ci/mmol; Perkin Elmer). Labeled pRNA was incubated with increasing amounts (2.5–160 nM) of recombinant proteins in RNA binding buffer (10 mM Tris–HCl [pH 7.4], 100 mM KCl, 5 mM $MgCl_2$) for 30 min on ice and filtered through nitrocellulose membranes (Protran, Whatman). To measure input RNA probes nylon membrane (GE Healthcare) was used. For measuring the specific binding affinity of TIP5/TAM-AT to pRNA (Figure 2d) 40 nM of TIP5/TAM-AT was first incubated with four times excess (2.5–80 nM) of unlabeled MCS-RNA for 30 min on ice. Then, increasing amounts (0.625–20 nM) of [^{32}P]-labeled pRNA probe were added and reactions were allowed to stay on ice for additional 1.5 h. After four washes with 10 volumes of RNA binding buffer, retained RNA was monitored by scintillation counting or by PhosphorImaging and quantified with Image Gauge 4.0 quantification software.

RNA immunoprecipitation assays (RIP)

HEK293T cells were transfected with expression vectors encoding FLAG/HA-tagged wild-type or mutant full-length human TIP5 or a control vector (pCS2+). Forty-eight hours after transfection cells were lysed in 500 μl of RNA immunoprecipitation assays (RIP) lysis buffer (200 mM NaCl, 5 mM $MgCl_2$, 10 mM Tris–HCl [pH 7.2], 0.5% NP-40, 1 mM DTT) supplemented with 100 u/ml of RNasin (Promega) and 25 μl/ml of proteinase inhibitor cocktail (Sigma). Lysates were sonicated, spun and digested with 30 u/100 μl of Dnase I (Roche) for 20 min at room temperature. After addition of 5 mM EDTA, FLAG-tagged proteins were bound to M2 beads (Sigma), washed 10× with 10 volumes of RIP lysis buffer. RNA was recovered with Trizol (Invitrogen) and subjected to RT-PCR using the forward primer 5′-CGATGGTGGCGTTTTT (hrDNA/-150/-130F) and the reverse primer 5′-CTCGGAGCGAAAGATATACC (hrDNA/-30/-50R).

RESULTS

The TAM domain of TIP5 adopts an extended MBD fold

Sequence alignments of the TAM domain of human TIP5 (BAZ2A) with phylogenetically related homologs (14,47–48) and canonical methyl-CpG binding domains (MBDs) show sequence conservation of the TAM domain in vertebrates, whereas sequence homology with canonical MBDs is less obvious (Figure 1b). Notably, the TAM domain comprises a unique 36 amino acid insert, which is not present in canonical MBDs and has two conserved AT-hooks adjacent to the TAM domain. The phylogenetically conserved region of the TAM domain extends at both the N- and C-terminal ends beyond the borders of the MBD identified by domain predictors, e.g. SMART (49,50) and Pfam (51).

To study the structure and RNA binding properties of the TAM domain, we analyzed TIP5/TAM-AT, a TIP5 fragment comprising residues 496–664, by NMR chemical shifts and ^{15}N relaxation experiments (Figure 2a). The ^{13}C secondary chemical shifts, ^{15}N relaxation rates and low $\{^1H\}$-^{15}N heteronuclear NOE values for the amides of residues 496–519 and 622–664 (comprising the AT-hooks) indicate that these regions are flexible and intrinsically disordered, while residues 520–621 are structured and define the globular fold of the TAM domain.

To identify the minimal RNA binding region of TIP5 we performed EMSA using TIP5 versions comprising residues 332–723 ('TIP5'), 496–664 ('TIP5/TAM-AT') and 516–623 ('TIP5/TAM') and ^{32}P-labeled MCS-RNA. As shown in Figure 2b and c all three TIP5 fragments bind to the RNA. While at 30 nM protein TIP5 quantitatively forms protein–RNA complexes and TIP5/TAM-AT retains 97% of the RNA binding activity, TIP5/TAM has only 75% of the

efficiency of TIP5 for RNA binding. These results indicate that the TAM-AT domain comprises the RNA binding region of TIP5 and suggest that the TAM domain is the principal RNA binding domain of TIP5. The presence of super-shifted protein–RNA products formed both by TIP5 and TIP5/TAM-AT, but not by TIP5/TAM, as well as a slight increase in RNA binding affinity observed with TIP5/TAM-AT as compared to TIP5/TAM, suggests that additional auxiliary RNA interaction may be contributed by the AT-hooks. We used FBA to quantitatively determine the binding affinity of TIP5/TAM-AT to pRNA. These experiments show that TIP5/TAM-AT binds to pRNA with a dissociation constant of $K_D = 5.5$ nM (Figure 2d).

Next, we determined the solution structure of TIP5/TAM by heteronuclear NMR spectroscopy. The secondary structure of the TAM domain is indicated by ^{13}C secondary chemical shifts and supported by H/D exchange experiments (Figure 2a, Supplementary Figure S1a). The ensemble of 20 lowest energy structures is well-defined by the NMR data with a backbone coordinate RMSD of 0.38 ± 0.08 Å and good structural statistics (Supplementary Figure S1b, Table 1). The TAM domain adopts an α/β fold, with a twisted β-sheet comprised of five antiparallel β-strands that pack against three α-helices (Figure 3a). The β-sheet represents a continuous surface, where two short loops connecting β1–β2 and β4–β5 open up like a clamp in the C-terminal region. Helix α2 (Phe563-Arg572) is inserted between the strands β3 and β4 and followed by a 3_{10}-helical turn. Helix α3 (Ala611-Ile621) follows the β5 strand at the C-terminus of the protein. The N-terminal helix α1 (Pro522-Gln530) is slightly bent, presumably due to the presence of a conserved proline residue (Pro528). The TAM domain adopts a stable and rigid fold, with only few residues in the linker following helix α2 (Asn573-Arg579), and the β1–β2 with β4–β5 loops showing increased flexibility as indicated by low $\{^1H\}$-^{15}N heteronuclear NOE values (Figure 2a; Supplementary Figure S1a). Notably, the linker connecting the 3_{10}-helical turn with β4 (Phe583-Gly592) is well-ordered and not flexible.

Although the structural core of the TAM domain resembles the MBD domain (Figure 3b) it exhibits unique N- and C-terminal extensions. Compared to the MBD fold the TAM domain has an additional N-terminal helix (α1), a small α/β motif comprising β4, β5 and a helix α3 at the C-terminus. The β1–β2 loop in the TAM domain corresponds to the L1 loop in MBDs which normally mediates contacts with the DNA backbone (52–54). However, the β1–β2 loop in the TAM domain is much shorter and adopts a different conformation, suggesting that this specific structural element may not contact methylated DNA and has a distinct role. Structure-based sequence alignments show that the unique structural elements of the TAM domain, i.e. specifically the small α/β motif, correspond to unique sequence insertions (Figure 1b) that are not present in canonical MBD domains.

Unique structural elements mediate the interaction of TIP5/TAM with RNA

To map the RNA binding surface of the TIP5/TAM-AT domain by NMR chemical shift perturbation experiments, we designed a shorter version of pRNA, dubbed $pRNA^{mini}$. $pRNA^{mini}$ has been previously described and represents a fusion of pRNA nucleotides -127/-113 with nucleotides -60/-49. This preserves a specific stem-loop structure that is recognized by TIP5 with features conserved between mouse and human (10,55) (Figure 4a, Supplementary Figure S2). The secondary structure and base pairing of $pRNA^{mini}$ was confirmed by an imino NOESY spectrum (Figure 4b). In competition EMSA experiments we found that pRNA and $pRNA^{mini}$ are able to compete binding of non-specific MCS RNA to TIP5 at low and high nanomolar concentrations, respectively (Supplementary Figure S2). This is consistent with the low nanomolar affinity observed for the pRNA-TIP5 interaction in EMSA experiments (Figure 2d). It also indicates that $pRNA^{mini}$ binds to TIP5 with a (high) nanomolar dissociation constant. Thus, although $pRNA^{mini}$ has (about 10-fold) reduced affinity to TIP5, our data suggest that it harbors important structural features required for efficient binding to TIP5 and was therefore used for NMR titration experiments.

We monitored binding of TIP5/TAM-AT to $pRNA^{mini}$ in a series of ^1H,^{15}N NMR correlation spectra with ^{15}N labeled, 50% deuterated TIP5/TAM-AT protein at increasing molar equivalents of RNA. Numerous NMR signals experience large chemical shift changes and/or line-broadening upon addition of $pRNA^{mini}$ (Figure 4c and d). The binding kinetics is in the fast to intermediate exchange regime with respect to the NMR chemical shift timescale, consistent with a dissociation constant in a high nanomolar range (56) as indicated in the competition EMSA experiments (Supplementary Figure S2). Interestingly, binding is saturated at 1:0.5 molar ratio of protein to RNA, suggesting that the RNA has two binding sites for TIP5/TAM-AT. The RNA binding surface of the TAM domain maps to the small α/β motif comprising β4, β5 and helix α3, i.e. the region that is unique to the TAM domain (Figure 3c). This RNA binding surface is flanked by a positively charged region, indicating that electrostatic interactions contribute to RNA binding. (Figure 3d). The AT-hooks harbor a number of additional positively charged residues. In the NMR titrations, many residues of the AT-hooks are strongly affected by RNA addition, although the chemical shift changes could in most cases not be traced to quantitatively analyze the chemical shift perturbations. Nevertheless, this indicates that the AT-hooks also contribute to the RNA binding, presumably by mediating additional electrostatic contacts with the RNA. In the free protein the AT-hooks are intrinsically disordered (Figure 2a), while in the RNA complex they may become more ordered by forming charged interactions with the RNA. This is supported by line broadening and/or large chemical shift changes for the few residues that could be analyzed in the NMR titration (i.e. A634, T636; Figure 4c and d). Together, the TAM domain harbors a novel RNA binding surface that involves a unique C-terminal α/β motif that is absent in canonical MBDs, thus rationalizing the unique capability of TAM to bind RNA. The AT-hooks

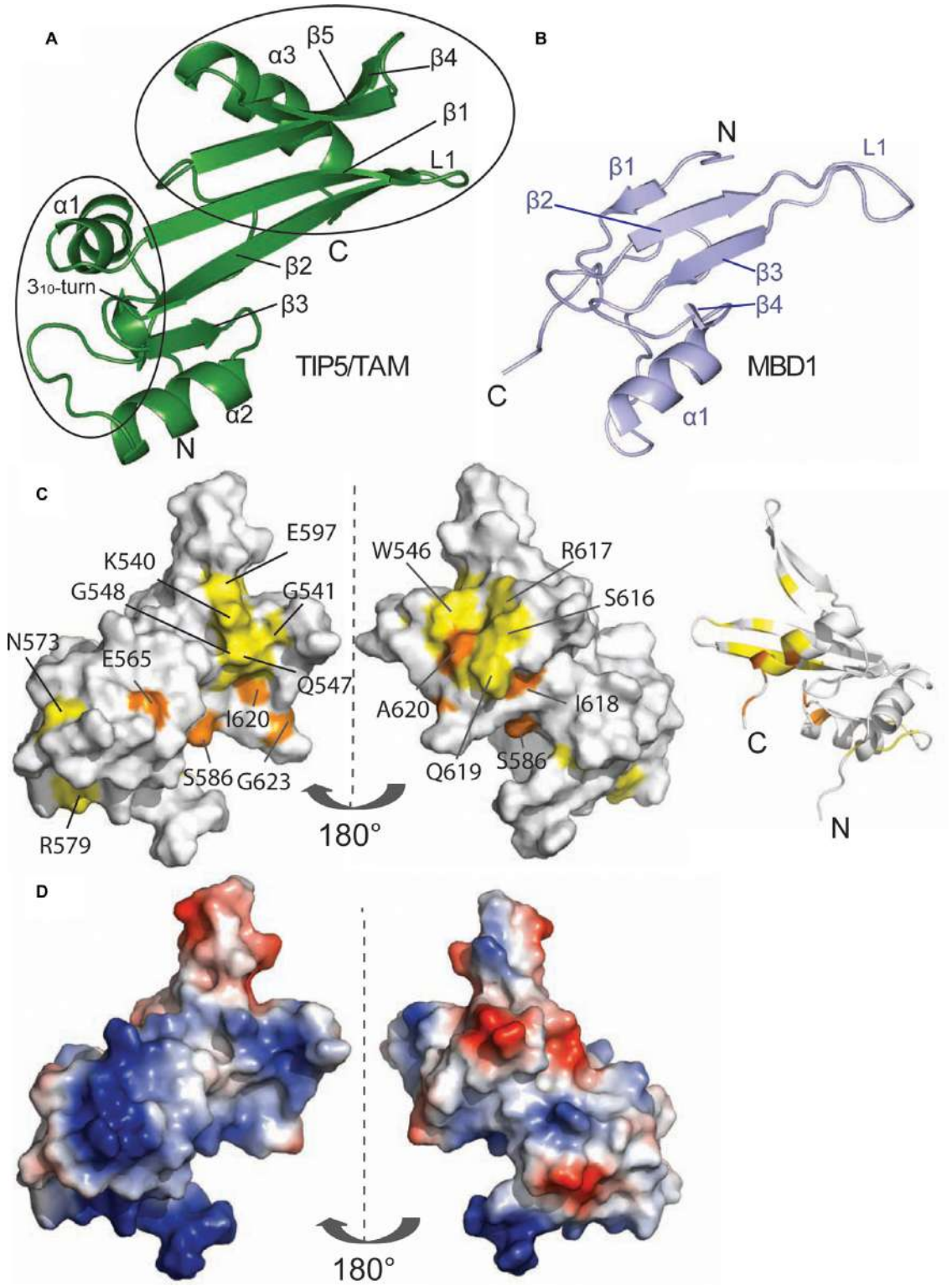


Figure 3. Solution structure of the hTIP5/TAM domain. (**A**) Cartoon representation of the human TIP5/TAM domain. Secondary structure elements and important loops are annotated. The TAM-specific N-terminal helix and the C-terminal α/β motif are highlighted by black circles. (**B**) Cartoon representation of the canonical MBD domain of human MBD1 (light blue, PDB ID: 1D9N) in a comparable view as the TAM domain on the left. Secondary structure elements and the loop L1 that mediates DNA binding of the MBD1 domain are annotated. (**C**) Mapping of the binding surface of TIP5/TAM with $pRNA^{mini}$ based on NMR chemical shift perturbations (Figure 4c) onto a surface representation (left and middle panel) and a ribbon model (right) of the TIP5/TAM domain structure. Left: same view as in (a), middle and right views are rotated by 180°. Colors indicate the RNA binding surface mapped by NMR titrations (Figure 4c and d). Residues with significant chemical shift perturbation are shown in yellow, while residues that exhibit strong intensity reductions upon RNA binding are colored orange as in Figure 4c and d. (**D**) Electrostatic potential of the TAM domain. Vacuum electrostatics surfaces are shown blue for positive and red for negative charges for the TAM domain in the same orientations as in (c).

Table 1. Structural statistics for the human TIP5/TAM domain

Number of NMR restraints	
NOE-derived distance restraints[a]	
Total NOE	2732
Intra/inter-residue	607/2125
Sequential ($\|i-j\|=1$)	738
Medium-range ($1<\|i-j\|<5$)	478
Long-range ($\|i-j\|\geq 5$)	909
Dihedral angle restraints (ϕ/ψ) [b]	83/85
Residual dipolar couplings[c]	178
Restraint violations (mean and s.d.)	
Distance restraints (Å)	0.013 ± 0.001
Dihedral angle restraints (°)	1.150 ± 0.046
Maximum dihedral angle violation (°)	4.96
Maximum distance constraint violation (Å)	0.28
RDC Q-factor (%)[c]	17.1%
Deviations from idealized geometry	
Bond lengths (Å)	0.004 ± 0.00
Bond angles (°)	0.536 ± 0.01
Impropers (°)	1.232 ± 0.05
Coordinate r.m.s. deviation (Å)	
Heavy	0.98 ± 0.13
Backbone	0.32 ± 0.07

[a]Distance restraints were employed with a soft square well potential using an energy constant of 50kcal mol^{-1} $Å^{-2}$.
[b]Torsion angle restraints derived from $TALOS_+$ (31) were applied to φ, ψ backbone angles using energy constants of 200 kcal $mol^{-1}rad^{-2}$.
[c]Residual dipolar couplings (RDCs) were employed with a harmonic potential using an energy constant of 0.8 kcal $mol^{-1}Hz^{-2}$ for the $^{1}H^{N}$-^{15}N RDCs and 0.4 kcal $mol^{-1}Hz^{-2}$ for the $^{13}C'$-^{15}N RDCs. Q-factor as defined by (46).
Structural statistics are given for the 20 lowest-energy structures after water refinement out of 200 calculated for residues 516–624 of the human TIP5/TAM domain. Pair wise coordinate r.m.s. deviations are calculated for residues 522–621. The CNS E_{repel} function was used to simulate van der Waals interactions with an energy constant of 25kcal mol^{-1} $Å^{-4}$ using PROLSQ van der Waals radii (41). Backbone torsion angles for 92.1, 7.8 and 0.1% of all residues are found in the core, allowed and generally allowed regions of the Ramachandran plot, none in disallowed regions.

likely contribute additional non-specific electrostatic contacts with the RNA.

We also monitored the interaction of TIP5/TAM with other RNAs (Supplementary Figure S3). While negligible chemical shift changes were observed with single-stranded poly-U RNA, significant spectral changes occurred in the presence of an RNA hairpin found in the A-repeat region of Xist a noncoding RNA that is required for X-chromosome inactivation (57,58). Interestingly, the TAM residues affected in NMR titrations upon binding to $pRNA^{mini}$ or Xist RNA are similar (Supplementary Figure S3c). This observation is consistent with the EMSA data and previous studies showing that TIP5 does not recognize a specific RNA sequence but rather distinct structural features, e.g. double-helical regions in stem-loop structures that may also be present in other RNAs.

Mutational analysis of the TAM–RNA interface

To validate the NMR data and examine the functional significance of the RNA binding surface, we monitored RNA binding of a panel of mutants that comprise specific amino acid exchanges in the TAM domain. Solvent exposed amino acid residues within the pRNA binding interface as well as conserved residues within the β-sheet surface of MBD domains (52) were altered, yielding the TAM variants K541E, K540E/K541E, R545E, W546Y, W546A, Q547A, Q562A, Q605A, S616D, R617E and Q619A (Figure 5a). The structural integrity of all proteins was monitored by NMR (Supplementary Figure S4). We also investigated the W551/Y552 double mutation in human TIP5, as we have previously shown that the corresponding mouse TIP5 W531G/Y532A mutant abolishes pRNA binding and strongly affects heterochromatic hallmarks when overexpressed in eukaryotic cells (6). The corresponding residues in human TIP5 (W551 and Y552) contribute to the hydrophobic core of the TAM domain fold. Consistent with this the double mutation W551G/Y552A is structurally unstable as the recombinant TAM W551G/Y552A double mutant is partially unfolded (Supplementary Figure S5). This shows that the structural integrity of the TAM domain is required for pRNA binding and functional activity of TIP5.

Next, we monitored RNA binding of TIP5, TIP5/TAM-AT and several TIP5/TAM-AT mutants using filter binding assays and EMSA. As expected, mutations altering residues remote from the TAM-pRNA interface (Q547A, Q562A, Q605A, Q619A and S616D) did not significantly affect RNA binding without affecting the tertiary structure (Supplementary Figure S4 and data not shown). In contrast, mutations within the RNA binding interface (K541E, W546Y, W546A and R617E) significantly reduced binding to pRNA, without affecting the domain fold (Figure 5b, Supplementary Figure 4). NMR spectra data indicate that the TAM domain fold is destabilized in the K540E/K541E double mutant. Thus, impaired RNA binding of this mutant is likely due to misfolding and/or aggregation of the TAM domain, underlining the importance of structural integrity for RNA binding of TIP5/TAM. Consistent with the filter binding experiments, electrophoretic shift mobility assays revealed that RNA binding of mutants R545E and W546A was strongly decreased (Figure 5c), indicating that R545 and W546 are directly involved in the interaction of TIP5 with RNA.

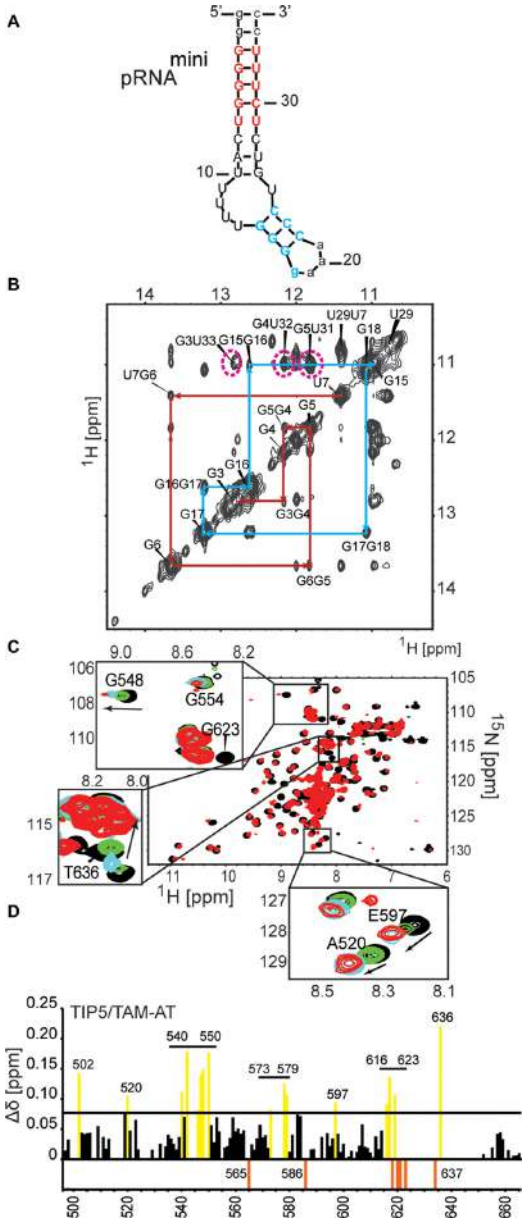


Figure 4. pRNA binding of the TAM domain. (**A**) pRNA^{mini} preserves the structural features of the pRNA stem recognized by TIP5. The secondary structure was calculated by Mfold (55). Nucleotides added for stabilization of the construct are shown in lower-case letters. (**B**) Confirmation of the secondary structure of pRNA^{mini} by an imino NOESY NMR spectrum. The sequential connectivities between imino protons of neighboring base pairs in the two stem regions are indicated by red and blue lines, respectively, matching the colors indicated for the RNA stem in (a). The characteristic imino signals of the G-U and U-U base pairs have chemical shifts around 11–12 ppm as expected. G-U cross peaks are indicated by magenta dotted lines. The G3-U33 cross peak is broadened possibly due to its terminal position in the pRNA^{mini} stem. (**C**) Superposition of ^{1}H,^{15}N HSQC NMR spectra of ^{15}N-labeled, 50% random fractional deuterated TIP5/TAM-AT free (black) and in the presence of a 1.2 molar excess of pRNA^{mini} (red). Insets show close-up views for specific residues upon addition of 0.2 (green), 0.5 (blue) and 1.2 (red) molar equivalents of pRNA^{mini}. Arrows indicate the direction of the peak shift. (**D**) NMR chemical shift perturbations observed for backbone amides in TIP5/TAM-AT at 1.2 molar excess of pRNA. NMR signals that experience line-broadening upon titration of the RNA are shown as orange negative bars. Residues with chemical shift changes larger than one standard deviation from the average ($\Delta\delta = 0.078$ ppm) are colored yellow.

To examine whether mutations that impair pRNA binding *in vitro* also affect TIP5-pRNA interaction *in vivo*, we performed RIP assays to monitor the association of pRNA with wild-type and mutant TIP5. Lysates from cells expressing FLAG-tagged TIP5, TIP5-R545E or TIP5-W546A were precipitated with anti-FLAG antibodies, and co-precipitated pRNA was monitored by RT-PCR. Cells transfected with the empty vector were used as a control. As shown in Figure 5d, RNA binding of the R545E or W546A mutant was markedly decreased as compared to wildtype TIP5. These results are in accord with the *in vitro* data demonstrating the functional relevance of residues R545 and W546 for TIP5 binding to RNA.

DISCUSSION

We report the solution structure of the TAM (TIP5/ARBP/MBD) domain of human TIP5 and show that it represents a novel extended MBD fold that is normally found in MBD-containing proteins such as MeCP1, MBD1, MBD2, MBD3 and MBD4. MBD-containing proteins function by recruiting various protein complexes to methyl-CpG sites, linking DNA methylation with chromatin modification and transcriptional repression. Significantly, the structurally related TAM domain does not exhibit binding preference to methylated DNA but associates with RNA. Previous studies have shown that the interaction of TIP5 with pRNA is crucial for targeting the chromatin remodeling complex NoRC to nucleoli and the establishment of heterochromatic features at a subset of rRNA genes (6). TIP5 binds to a phylogenetically conserved hairpin structure of pRNA, which is required for epigenetic regulation of rRNA genes (10).

Our structure reveals that the structural core of the TAM domain resembles the MBDs of MBD1 and MeCP2. However, the TAM domain comprises unique C-terminal extensions, with the C-terminal α/β motif not being present in canonical MBD structures. Significantly, the positively charged C-terminal α/β motif constitutes the binding interface of TIP5 and pRNA. The sequence of the TAM domain and residues involved in RNA binding are conserved in vertebrates, underscoring the functional importance of the TAM domain distinct from MBD proteins. Mutational analysis revealed that residues in the RNA binding surface are required for RNA binding *in vitro* and *in vivo*. We found that two of the tested residues, W546 and R545, are the most important for RNA binding. Notably, these residues are not conserved in human MBDs, suggesting that the interaction of TAM-containing proteins with specific regulatory RNA may guide chromatin modifying enzymes to regulatory gene regions.

Canonical MBDs recognize methylated CpGs in DNA by hydrophobic residues located at the β-sheet surface, with additional DNA backbone contacts mediated by the extended loop L1 (52) (Figure 6). Although the fold of the TAM domain resembles the MBD domain, TAM does not exhibit binding preference for methylated DNA (Supplementary Figure S6). This may be explained by the fact that amino acids that mediate specific recognition of methylated CpG regions by the MBD domain are not conserved in the TAM domain (Figure 1b). Hence, the TIP5/TAM domain

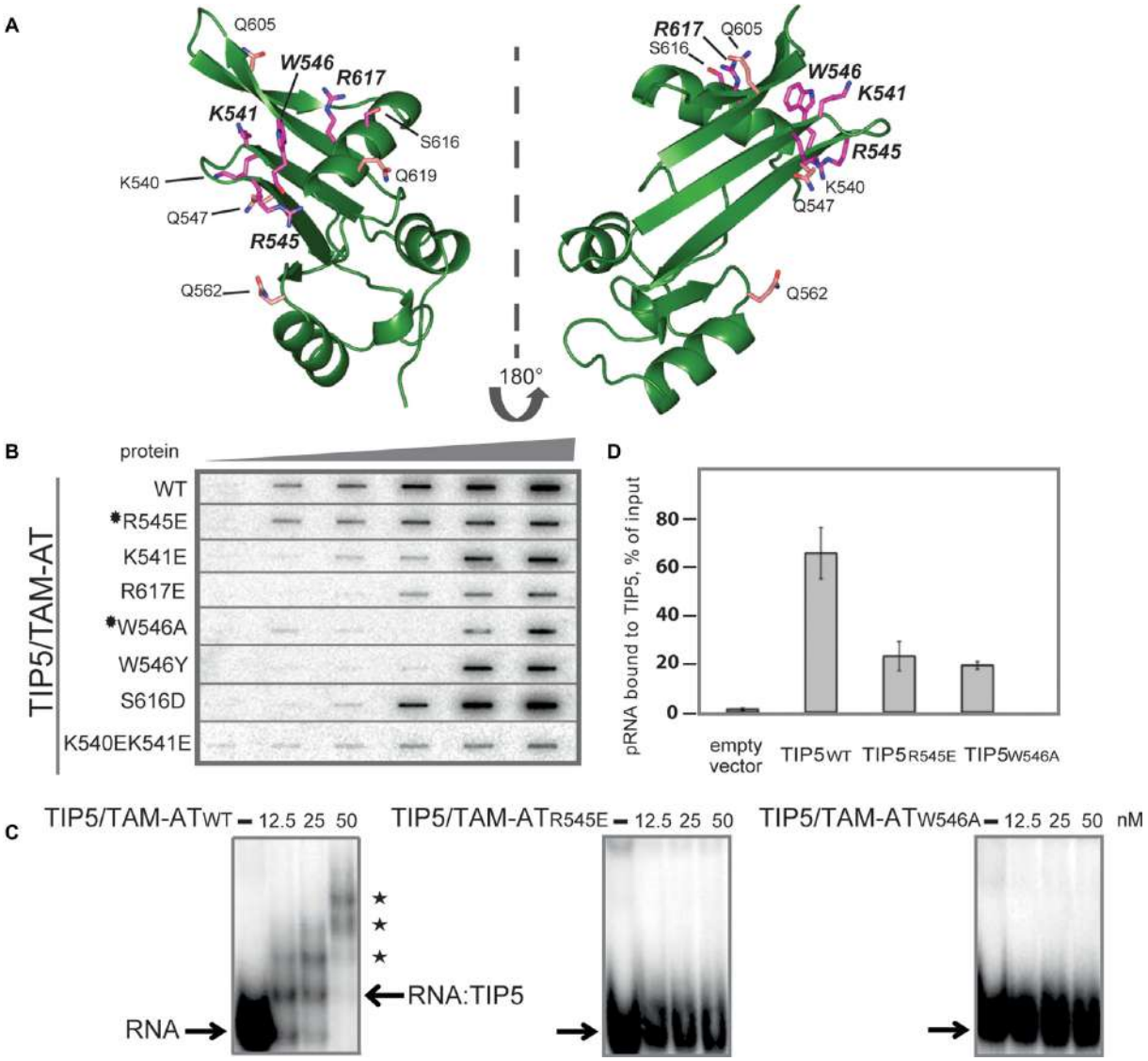


Figure 5. Mutational analysis of TAM/pRNA interaction. (**A**) Side chains of residues probed by mutational analysis are shown in magenta on the structure of the TIP5/TAM domain. Residues which are involved in RNA binding are annotated with larger font. Side chains shown in pink are controls, not expected to be involved in RNA binding. (**B**) RNA binding of wild-type and mutant TIP5/TAM-AT. The triangle on top indicates increasing protein concentrations (0, 5, 10, 20, 40 and 80 nM) of wild-type TIP5/TAM-AT (WT) or the indicated TIP5/TAM-AT mutants, which were incubated with 10 nM of [^{32}P]-pRNA. Protein–RNA complexes were retained on nitrocellulose filters and visualized by phosphorimaging. Mutants selected for further analyses (R545E and W546A) are marked with black stars. (**C**) EMSA experiments monitoring binding of wild-type TIP5/TAM-AT and mutants R545E and W546A to RNA. 12.5, 25, 50 nM of the respective proteins were incubated with 2.5 nM of [^{32}P]-labeled RNA. Positions of free RNA and RNA:TIP5 complexes are indicated by arrows. Supershifts indicating higher order complexes of TIP5 binding with several RNA molecules are marked with asterisks. (**D**) RNA immunoprecipitation experiments monitoring binding of wild-type and mutant TIP5 to pRNA *in vivo*. HEK293T cells were transfected with expression vectors encoding FLAG-tagged wild-type TIP5, TIP5/R545E or TIP5/W546A. TIP5 was immunoprecipitated with anti-FLAG antibodies and associated pRNA was monitored by RT-qPCR.

has evolved as a specific pRNA recognition domain that is required for NoRC recruitment to the rDNA promoter.

Interestingly, our NMR titrations show that the TAM domain also binds to a hairpin found in Xist RNA, a non-coding RNA involved in dosage compensation and X-chromosome inactivation. A common structural feature between pRNA and Xist RNA are double-helical regions. This suggests that the TIP5/TAM domain preferably recognizes double-helical regions in NoRC-associated RNAs. This may explain why our NMR titrations indicate that two TAM-AT proteins can bind to pRNA^{mini}, which has two helical stem regions. Also, the observation of supershifts in the EMSA experiments demonstrates the formation of higher order complexes, which could involve binding of TIP5 to different helical stem regions in larger RNA ligands.

A general preference for helical regions in RNAs by TIP5 is supported by recent data showing that the function of NoRC is not restricted to silencing of rRNA genes but is also required for heterochromatin formation and silencing of centromeres and telomeres (12). Like rRNA genes,

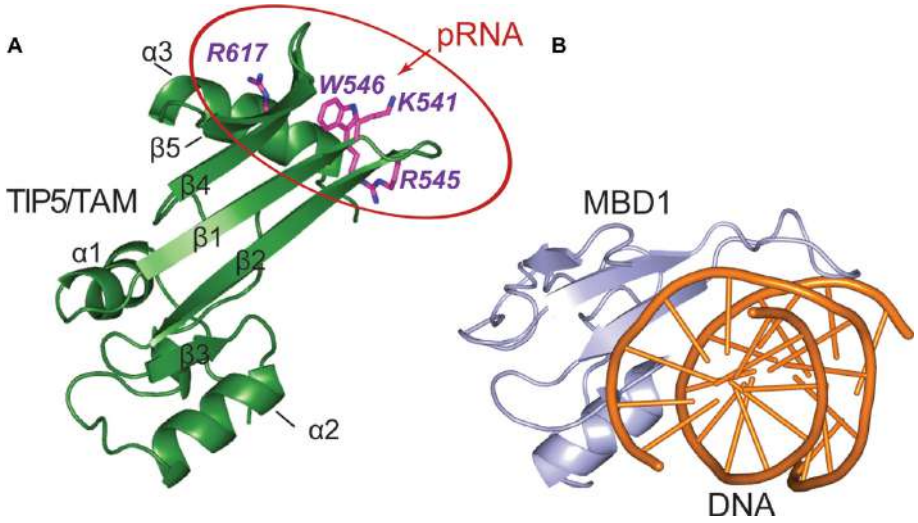


Figure 6. Comparison of TAM and MBD nucleic acid binding domains. (**A**) The pRNA binding surface of the TIP5/TAM domain. Side chains of residues important for RNA binding are shown on a cartoon representation of the TIP5/TAM domain structure. (**B**) Cartoon representation of the canonical MBD domain from MBD1 (light blue) in complex with methylated DNA shown as orange cartoon (PDB ID: 1IG4) (52).

centromeres and telomeres are organized in repeated clusters and exhibit a repressive heterochromatic structure. This heterochromatic structure depends on specific ncRNAs, i.e. satellite transcripts originating from the centromeric core region and on TERRA ('telomeric repeat-containing RNA'), respectively. Indeed, TERRA was shown to be associated with TIP5, supporting that TERRA serves a guiding function in NoRC-dependent heterochromatin formation at telomeres (12). It is tempting to speculate that the interaction of NoRC with pRNA, satellite RNA and TERRA relies on the recognition of double-helical regions in these RNAs which is recognized by the TAM domain of TIP5. Future structural characterization of full-length TIP5 and its interaction with different RNAs will give further insights into the versatility of RNA recognition motifs and their involvement in diverse biological functions.

ACCESSION NUMBERS

The atomic coordinates for the NMR ensemble of the TIP5/TAM domain have been deposited in the Protein Data Bank under accession number 5AGQ. The chemical shift assignments have been deposited in the Biological Magnetic Resonance Data Bank under accession number 26517.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors are grateful to Christine Mayer for initiating this study, the design of pRNA^{mini} and fruitful discussions. *Author Contributions*: I.A. performed molecular biology, NMR analysis and structure calculation, I.A. and S.M. performed protein purification, K.T. and F.K. contributed to NMR analysis and structure calculation, S.M. and C.M. performed RNA binding experiments. M.S. and I.G. conceived the project. I.A., S.M., I.G. and M.S wrote the paper. All authors discussed the results and commented on the manuscript.

FUNDING

Alexander von Humboldt foundation (to K.T., F.K.); Deutsche Forschungsgemeinschaft [SFB1035 to M.S.; GR475/22-1, SFB1036 to I.G.]; Excellence Clusters Center for integrated Protein Science Munich (to M.S.); CellNetworks (to I.G.).
Conflict of interest statement. None declared.

REFERENCES

1. McStay,B. and Grummt,I. (2008) The epigenetics of rRNA genes: from molecular to chromosome biology. *Annu. Rev. Cell Dev. Biol.*, **24**, 131–157.
2. Strohner,R., Nemeth,A., Jansa,P., Hofmann-Rohrer,U., Santoro,R., Langst,G. and Grummt,I. (2001) NoRC–a novel member of mammalian ISWI-containing chromatin remodeling machines. *Embo J.*, **20**, 4892–4900.
3. Santoro,R., Li,J. and Grummt,I. (2002) The nucleolar remodeling complex NoRC mediates heterochromatin formation and silencing of ribosomal gene transcription. *Nat. Genet.*, **32**, 393–396.
4. Zhou,Y., Santoro,R. and Grummt,I. (2002) The chromatin remodeling complex NoRC targets HDAC1 to the ribosomal gene promoter and represses RNA polymerase I transcription. *EMBO J.*, **21**, 4632–4640.
5. Li,J., Langst,G. and Grummt,I. (2006) NoRC-dependent nucleosome positioning silences rRNA genes. *EMBO J.*, **25**, 5735–5741.
6. Mayer,C., Schmitz,K.M., Li,J., Grummt,I. and Santoro,R. (2006) Intergenic transcripts regulate the epigenetic state of rRNA genes. *Mol. Cell*, **22**, 351–361.
7. Schmitz,K.M., Mayer,C., Postepska,A. and Grummt,I. (2010) Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev.*, **24**, 2264–2269.
8. Zhou,Y. and Grummt,I. (2005) The PHD finger/bromodomain of NoRC interacts with acetylated histone H4K16 and is sufficient for rDNA silencing. *Curr. Biol.*, **15**, 1434–1438.
9. Santoro,R., Schmitz,K.M., Sandoval,J. and Grummt,I. (2010) Intergenic transcripts originating from a subclass of ribosomal DNA

repeats silence ribosomal RNA genes in trans. *EMBO Rep.*, **11**, 52–58.
10. Mayer,C., Neubert,M. and Grummt,I. (2008) The structure of NoRC-associated RNA is crucial for targeting the chromatin remodelling complex NoRC to the nucleolus. *EMBO Rep.*, **9**, 774–780.
11. Savic,N., Bar,D., Leone,S., Frommel,S.C., Weber,F.A., Vollenweider,E., Ferrari,E., Ziegler,U., Kaech,A., Shakhova,O. *et al.* (2014) lncRNA maturation to initiate heterochromatin formation in the nucleolus is required for exit from pluripotency in ESCs. *Cell Stem Cell*, **15**, 720–734.
12. Postepska-Igielska,A., Krunic,D., Schmitt,N., Greulich-Bode,K.M., Boukamp,P. and Grummt,I. (2013) The chromatin remodelling complex NoRC safeguards genome stability by heterochromatin formation at telomeres and centromeres. *EMBO Rep.*, **14**, 704–710.
13. Gu,L., Frommel,S.C., Oakes,C.C., Simon,R., Grupp,K., Gerig,C.Y., Bar,D., Robinson,M.D., Baer,C., Weiss,M. *et al.* (2015) BAZ2A (TIP5) is involved in epigenetic alterations in prostate cancer and its overexpression predicts disease recurrence. *Nat. Genet.*, **47**, 22–30.
14. Hendrich,B. and Tweedie,S. (2003) The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet.*, **19**, 269–277.
15. Aravind,L. and Landsman,D. (1998) AT-hook motifs identified in a wide variety of DNA-binding proteins. *Nucleic Acids Res.*, **26**, 4413–4421.
16. Jeffery,L. and Nakielny,S. (2004) Components of the DNA methylation system of chromatin control are RNA-binding proteins. *J. Biol. Chem.*, **279**, 49479–49487.
17. Lukavsky,P.J. and Puglisi,J.D. (2004) Large-scale preparation and purification of polyacrylamide-free RNA oligonucleotides. *RNA*, **10**, 889–893.
18. Wyatt,J.R., Chastain,M. and Puglisi,J.D. (1991) Synthesis and purification of large amounts of RNA oligonucleotides. *Biotechniques*, **11**, 764–769.
19. Kao,C., Zheng,M. and Rudisser,S. (1999) A simple and efficient method to reduce nontemplated nucleotide addition at the 3 terminus of RNAs transcribed by T7 RNA polymerase. *RNA*, **5**, 1268–1272.
20. Delaglio,F., Grzesiek,S., Vuister,G.W., Zhu,G., Pfeifer,J. and Bax,A. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR*, **6**, 277–293.
21. Sattler,M., Schleucher,J. and Griesinger,C. (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nucl. Magn. Reson. Spectrosc.*, **34**, 93–158.
22. Pastore,A. and Saudek,V. (1990) The relationship between chemical shift and secondary structure in proteins. *J. Magn. Reson.*, **90**, 165–176.
23. Spera,S. and Bax,A. (1991) Empirical correlation between protein backbone conformation and C.alpha. and C.beta. 13C nuclear magnetic resonance chemical shifts. *J. Am. Chem. Soc.*, **113**, 5490–5492.
24. Yamazaki,T., Forman-Kay,J.D. and Kay,L.E. (1993) Two-dimensional NMR experiments for correlating carbon-13.beta. and proton.delta./.epsilon. chemical shifts of aromatic residues in 13C-labeled proteins via scalar couplings. *J. Am. Chem. Soc.*, **115**, 11054–11055.
25. Pelton,J.G., Torchia,D.A., Meadow,N.D. and Roseman,S. (1993) Tautomeric states of the active-site histidines of phosphorylated and unphosphorylated IIIGlc, a signal-transducing protein from Escherichia coli, using two-dimensional heteronuclear NMR techniques. *Protein Sci.*, **2**, 543–558.
26. Senn,H., Werner,B., Messerle,B.A., Weber,C., Traber,R. and Wuethrich,K. (1989) Stereospecific assignment of the methyl 1H NMR lines of valine and leucine in polypeptides by nonrandom 13C labelling *FEBS Lett.*, **249**, 113–118.
27. Neri,D., Szyperski,T., Otting,G., Senn,H. and Wüthrich,K. (1989) Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 repressor by biosynthetically directed fractional ^{13}C labeling. *Biochemistry*, **28**, 7510–7516.
28. Jung,Y.S. and Zweckstetter,M. (2004) Mars—robust automatic backbone assignment of proteins. *J. Biomol. NMR*, **30**, 11–23.
29. Guntert,P. (2004) Automated NMR structure calculation with CYANA. *Methods Mol. Biol.*, **278**, 353–378.
30. Shen,Y., Delaglio,F., Cornilescu,G. and Bax,A. (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR*, **44**, 213–223.
31. Cordier,F., Dingley,A.J. and Grzesiek,S. (1999) A doublet-separated sensitivity-enhanced HSQC for the determination of scalar and dipolar one-bond J-couplings. *J. Biomol. NMR*, **13**, 175–180.
32. Ishii,Y., Markus,M.A. and Tycko,R. (2001) Controlling residual dipolar couplings in high-resolution NMR of proteins by strain induced alignment in a gel. *J. Biomol. NMR*, **21**, 141–151.
33. Ottiger,M. and Bax,A. (1999) Bicelle-based liquid crystals for NMR-measurement of dipolar couplings at acidic and basic pH values. *J. Biomol. NMR*, **13**, 187–191.
34. Farrow,N.A., Muhandiram,R., Singer,A.U., Pascal,S.M., Kay,C.M., Gish,G., Shoelson,S.E., Pawson,T., Forman-Kay,J.D. and Kay,L.E. (1994) Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by 15N NMR relaxation. *Biochemistry*, **33**, 5984–6003.
35. Akke,M. and Palmer,A.G. (1996) Monitoring macromolecular motions on microsecond to millisecond time scales by $R1\rho - R1$ constant relaxation time NMR spectroscopy. *J. Am. Chem. Soc.*, **118**, 911–912.
36. Kay,L.E., Torchia,D.A. and Bax,A. (1989) Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry*, **28**, 8972–8979.
37. Bruschweiler,R., Liao,X. and Wright,P. (1995) Long-range motional restrictions in a multidomain zinc-finger protein from anisotropic tumbling. *Science*, **268**, 886–889.
38. Daragan,V.A. and Mayo,K.H. (1997) Motional model analyses of protein and peptide dynamics using 13C and 15N NMR relaxation. *Prog. Nucl. Magn. Reson. Spectrosc.*, **31**, 63–105.
39. Schanda,P. and Brutscher,B. (2005) Very fast two-dimensional NMR spectroscopy for real-time investigation of dynamic events in proteins on the time scale of seconds. *J. Am. Chem. Soc.*, **127**, 8014–8015.
40. Linge,J.P., Habeck,M., Rieping,W. and Nilges,M. (2003) ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics*, **19**, 315–316.
41. Linge,J.P., Williams,M.A., Spronk,C.A., Bonvin,A.M. and Nilges,M. (2003) Refinement of protein structures in explicit solvent. *Proteins*, **50**, 496–506.
42. Doreleijers,J.F., Sousa da Silva,A.W., Krieger,E., Nabuurs,S.B., Spronk,C.A., Stevens,T.J., Vranken,W.F., Vriend,G. and Vuister,G.W. (2012) CING: an integrated residue-based structure validation program suite. *J. Biomol. NMR*, **54**, 267–283.
43. Laskowski,R.A., Macarthur,M.W., Moss,D.S. and Thornton,J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, **26**, 283–291.
44. Vriend,G. and Sander,C. (1993) Quality control of protein models: directional atomic contact analysis. *J. Appl. Crystallogr.*, **26**, 47–60.
45. Cornilescu,G., Marquardt,J.L., Ottiger,M. and Bax,A. (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J. Am. Chem. Soc.*, **120**, 6836–6837.
46. Stepanov,A.S., Voronina,A.S., Ovchinnikov,L.P. and Spirin,A.S. (1971) RNA-binding protein factor of animal cell extracts. *FEBS Lett.*, **18**, 13–18.
47. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
48. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
49. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.
50. Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 5857–5864.
51. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
52. Ohki,I., Shimotake,N., Fujita,N., Jee,J., Ikegami,T., Nakao,M. and Shirakawa,M. (2001) Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA. *Cell*, **105**, 487–497.

53. Ho,K.L., McNae,I.W., Schmiedeberg,L., Klose,R.J., Bird,A.P. and Walkinshaw,M.D. (2008) MeCP2 binding to DNA depends upon hydration at methyl-CpG. *Mol. Cell*, **29**, 525–531.
54. Scarsdale,J.N., Webb,H.D., Ginder,G.D. and Williams,D.C. Jr (2011) Solution structure and dynamic analysis of chicken MBD2 methyl binding domain bound to a target-methylated DNA sequence. *Nucleic Acids Res.*, **39**, 6741–6752.
55. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
56. Gobl,C., Madl,T., Simon,B. and Sattler,M. (2014) NMR approaches for structural analysis of multidomain proteins and complexes in solution. *Prog. Nucl. Magn. Reson. Spectrosc.*, **80C**, 26–63.
57. Postepska-Igielska,A. and Grummt,I. (2014) NoRC silences rRNA genes, telomeres, and centromeres. *Cell Cycle*, **13**, 493–494.
58. Bierhoff,H., Postepska-Igielska,A. and Grummt,I. (2014) Noisy silence: non-coding RNA and heterochromatin formation at repetitive elements. *Epigenetics*, **9**, 53–61.

THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 290, NO. 36, pp. 22111–22126, September 4, 2015
© 2015 by The American Society for Biochemistry and Molecular Biology, Inc. Published in the U.S.A.

A Crystallin Fold in the Interleukin-4-inducing Principle of *Schistosoma mansoni* Eggs (IPSE/α-1) Mediates IgE Binding for Antigen-independent Basophil Activation*

Received for publication, June 25, 2015 Published, JBC Papers in Press, July 10, 2015, DOI 10.1074/jbc.M115.675066

N. Helge Meyer[‡§1], Hubert Mayerhofer[¶], Konstantinos Tripsianes[‡§2], Silke Blindow[‖], Daniela Barths[‖], Astrid Mewes[‖], Thomas Weimar[**], Thies Köhli[**], Steffen Bade[‡‡], Tobias Madl[‡§3], Andreas Frey[‡‡], Helmut Haas[‖], Jochen Mueller-Dieckmann[¶4], Michael Sattler[‡§5], and Gabriele Schramm[‖6]

*From the [‡]Institute of Structural Biology, Helmholtz Zentrum München, 85764 Neuherberg, the [§]Munich Center for Integrated Protein Science and Biomolecular NMR, Department of Chemie, Technische Universität München, 85747 Garching, the [¶]EMBL Hamburg Outstation % Desy, 22603 Hamburg, the [‖]Cellular Allergology and [‡‡]Mucosal Immunology and Diagnostic Section, Research Center Borstel, 23845 Borstel, and the **Institute of Chemistry, University of Lübeck, 23562 Lübeck, Germany*

Background: The interleukin-4-inducing principle from *Schistosoma mansoni* eggs (IPSE/α-1) triggers basophils to release interleukin-4 and interleukin-13 in an IgE-dependent but antigen-independent way.
Results: Structural analysis identified IPSE/α-1 as a new member of the βγ-crystallin superfamily with a unique IgE-binding loop.
Conclusion: IPSE/α-1 activates basophils via IgE-binding crystallin folds.
Significance: Schistosomes use unique mechanisms to manipulate the host's immune response.

The IL-4-inducing principle from *Schistosoma mansoni* eggs (IPSE/α-1), the major secretory product of eggs from the parasitic worm *S. mansoni*, efficiently triggers basophils to release the immunomodulatory key cytokine interleukin-4. Activation by IPSE/α-1 requires the presence of IgE on the basophils, but the detailed molecular mechanism underlying activation is unknown. NMR and crystallographic analysis of IPSEΔNLS, a monomeric IPSE/α-1 mutant, revealed that IPSE/α-1 is a new member of the βγ-crystallin superfamily. We demonstrate that this molecule is a general immunoglobulin-binding factor with highest affinity for IgE. NMR binding studies of IPSEΔNLS with the 180-kDa molecule IgE identified a large positively charged binding surface that includes a flexible loop, which is unique to the IPSE/α-1 crystallin fold. Mutational analysis of amino acids in the binding interface showed that residues contributing to IgE binding are important for IgE-dependent activation of basophils. As IPSE/α-1 is unable to cross-link IgE, we propose that this molecule, by taking advantage of its unique IgE-binding crystallin fold, activates basophils by a novel, cross-linking-independent mechanism.

Schistosomes are parasitic worms causing a chronic disease with more than 200,000 deaths per year alone in sub-Saharan Africa (World Health Organization, Schistosomiasis Facts, 2012). During acute infection, schistosome eggs induce an inflammatory T helper type 2 (Th2) response characterized by high immunoglobulin (Ig) E titers and eosinophilia (1). Subsequently, the Th2 response is down-regulated by anti-inflammatory mechanisms to restrict host tissue damage (1). In this context, the cytokine interleukin (IL)-4 plays an important role, first as key inducer cytokine for a Th2 response (2) and second as a key factor for limiting extensive inflammation. The absence of IL-4 signaling in *Schistosoma mansoni*-infected mice, which are unable to produce IL-4 (IL-4$^{-/-}$ mice) or to react to IL-4 (IL-4 receptor α-chain$^{-/-}$ mice), is associated with massive intestinal inflammation and rapid death (3, 4). Recent studies suggest that basophils are involved in anti-helminth immunity (5, 6). They have been shown to migrate to parasite-affected tissues, where they become fully activated and represent the major source of IL-4. However, the mechanisms by which the parasites activate basophils remained elusive.

Previously, we identified a homodimeric 35-kDa glycoprotein secreted from live *S. mansoni* eggs, which induces the release of IL-4 from human and murine basophils (7, 8). This protein was called IPSE (IL-4-inducing principle from *S. man-*

* This work was supported in part by Deutsche Forschungsgemeinschaft Grants EXC114 and SFB 594 (to M. S) and the Bavarian Ministry of Sciences, Research, and the Arts in the framework of the Bavarian Molecular Biosystems Research Network (to T. M.). The authors declare that they have no conflicts of interest with the contents of this article.
The atomic coordinates and structure factors (codes 4AKA and 4EL6) have been deposited in the Protein Data Bank (http://wwpdb.org/).
Submitted to the BMRB Data Bank under accession number 17405.
[1] Present address: European Medical School, Dept. of General and Visceral Surgery, Universität Oldenburg, 26131 Oldenburg, Germany.
[2] Supported by the Alexander von Humboldt Foundation. Present address: Central European Institute of Technology, Masaryk University, 601 77Brno, Czech Republic.
[3] Supported by an EMBO Long Term Fellowship, a Schrödinger Fellowship by the Austrian Science Fund, and an APART Fellowship from the Austrian Academy of Sciences. Present address: Institute of Molecular Biology and Biochemistry, Center of Molecular Medicine, Medical University of Graz, 8010 Graz, Austria.
[4] To whom correspondence may be addressed. Tel.: 49-40-42816-660; E-mail: jochen.mueller-dieckmann@uni-hamburg.de.
[5] To whom correspondence may be addressed: Institute of Structural Biology, Helmholtz Zentrum München, 85764 Neuherberg, Munich Center for Integrated Protein Science and Biomolecular NMR, Dept. of Chemie, Technische Universität München, Lichtenbergstr. 4, 85747 Garching, Germany. Tel.: 49-89-289-13418; Fax: 49-89-289-13210, E-mail: Sattler@helmholtz-muenchen.de.
[6] To whom correspondence may be addressed: Cellular Allergology, Research Center Borstel, Parkallee 1–40, 23845 Borstel, Germany. Tel.: 49-4537-1885610; E-mail: gschramm@fz-borstel.de.

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

soni eggs) (7) but was subsequently renamed IPSE/α-1 (9) because it was found to be identical with the major excretory/secretory schistosome egg antigen α-1 described earlier (10). IPSE/α-1 has a unique sequence; apart from a predicted putative greek key motif, there is no similarity to any other known protein, except to translated cDNA sequences from *Schistosoma japonicum* and translated genomic sequences from *Schistosoma haematobium* (11). IPSE/α-1 triggers the release of IL-4 from basophils of nonsensitized healthy human donors, and intravenous injection of IPSE/α-1 into mice leads to basophil-IL-4 release in the liver, which, apart from the gut, is the major site of egg deposition during *S. mansoni* infection (8). Basophil activation by IPSE/α-1 was found to depend on the presence of IgE on the surface of the basophils irrespective of IgE's antigen specificity (7, 12). IgE-dependent, antigen-independent basophil activation is well known from multivalent lectins or B cell superantigens, which cross-link IgE via binding to its carbohydrate side chains or directly to its immunoglobulin backbone, respectively (13, 14). However, molecular details of the mechanism of action of IPSE/α-1 are not yet known.

Here, we report NMR and crystallographic structures of IPSEΔNLS, a monomeric deletion mutant of IPSE/α-1. IPSEΔNLS lacks a stretch of 10 amino acids at the C terminus comprising a putative nuclear localization sequence (NLS)[7] (15) and Cys-132, which mediates a disulfide bridge for dimer formation of the native molecule (16). The structure consists of two symmetrically arranged greek key motifs and exhibits high structural similarity with members of the βγ-crystallin superfamily. Biochemical studies show that IPSE/α-1 is a general immunoglobulin-binding factor binding to all isotypes but with highest affinity to IgE. NMR chemical shift analysis combined with functional assays suggest that IPSE/α-1 interacts with IgE via a large surface area that includes a loop comprising positively charged and aromatic amino acids, which are critical for IgE binding. In contrast to lectins and B cell superantigens, our data obtained with soluble or immobilized molecules suggest that IPSE/α-1 is not able to cross-link IgE. Thus, we propose that IPSE/α-1 activates basophils by a novel IgE-dependent, antigen-independent mechanism that does not involve cross-linking of IgE.

Experimental Procedures

Sample Preparation—Unlabeled IPSE/α-1 protein has been prepared as described elsewhere (17). For preparation of uniformly ^{13}C,^{15}N- and ^{15}N-labeled samples, bacteria were grown in M9 medium supplemented with ^{13}C-labeled glucose and/or $^{15}NH_4Cl$, respectively. For preparation of $^{2}H/^{15}N/^{13}C$-labeled IPSEΔNLS, the same protocol was employed but using [U-$^{2}H^{13}C$]glucose and growing bacteria in 100% D_2O. Point mutations were introduced in pProExHTb-IPSEΔNLS and pProExHTb IPSE using the QuikChange site-directed mutagenesis protocol (Invitrogen). The corresponding proteins contain two additional residues at the N terminus from a tobacco etch virus protease cleavage site, which was introduced by the expression vector.

NMR spectroscopy was carried out on a Bruker Avance III 750 MHz spectrometer equipped with a TXI probehead, a Bruker Avance III 600 MHz spectrometer equipped with a cryogenic TCI probehead, or on an Avance I 900 MHz spectrometer equipped with a cryogenic TXI probehead. Unless stated otherwise, all experiments were carried out on a 1 mM uniformly $^{13}C/^{15}N$-labeled IPSEΔNLS at 298 K. Spectra were processed with NMRPipe (18) and analyzed with Sparky3. For backbone and side chain assignment CBCA(CO)NH, CBCANH, and (H)CCH-TOCSY (19), spectra were recorded (20). Distance information was obtained from ^{15}N- and ^{13}C-edited NOESY spectra with a mixing time of 70 ms. ^{15}N R_1 and R_2 relaxation rates and $\{^{1}H\}$-^{15}N heteronuclear NOE data were measured at 750 MHz proton Larmor frequency and 298 K as described (21). Relaxation delays of 0.0216, 0.054, 0.27, 0.81, 1.08, 1.4, and 1.62 s were used to measure R_1 and 0.0144, 0.0288, 0.0576, 0.0864, 0.115, 0.173, and 0.173 were used to extract R_2 rates. Errors were estimated from duplicate measurements of two delays and Monte Carlo simulation. Uncertainty of the ratio of R_2/R_1 relaxation rates was determined by error propagation. Residue-specific tumbling correlation times, τ_c, were calculated as described elsewhere (21).

Structure calculation was performed by automated NOE cross-peak assignment using the software CYANA 3.0 (22). Automatically assigned NOEs and completeness of the NOE cross-peaks were inspected manually. Distance restraints from the CYANA calculation and TALOS+ (23)-derived torsion angles were used in a water refinement calculation (24) applying the RECOORD protocol (25). Ramachandran plot statistics after refinement were as follows: 91.1% of residues in most favored regions, 8.9% in additional allowed regions, 0% in generously allowed regions, and 0% in disallowed regions. Quality of the NMR-derived structure ensemble was validated using the iCING web server as well as PROCHECK (26) and WHATCHECK (27). Molecular images were generated using PyMOL (Delano Scientific).

NMR Titrations—Uniformly $^{2}H/^{15}N/^{13}C$-labeled IPSEΔNLS was stepwise added to 100 μM IgE (Dianova) at ratios of 0.5:1 and 1:1 (IPSEΔNLS/IgE), and chemical shifts were recorded using CRINEPT-HMQC spectra (28). A CRINEPT spectrum of 100 μM $^{2}H/^{15}N/^{13}C$-labeled IPSEΔNLS was recorded as a reference.

X-ray Structure Determination—IPSEΔNLS was solved by molecular replacement using Phaser as implemented in CCP4 6.1.131 with the modified NMR structure as a search model. Coordinates of IPSEΔNLS were refined using REFMAC 5.6.0119 alternating with rounds of manual model building using the program Coot. Initial crystallization trials were performed in 96-well sitting drop vapor diffusion plates at 20 °C at a protein concentration of 14 mg ml^{-1} using commercially available screens. Two conditions yielded small crystals, which were optimized through a combination of OptiSalt screening (Qiagen) and customized grid screens resulting in well diffracting crystals, which appeared after at least 1 week. Crystals were flash-frozen at 100 K in precipitant solution containing 15% glycerol. Diffraction data were collected at a wavelength of

[7] The abbreviations used are: NLS, nuclear localization sequence; r.m.s.d., root mean square deviation; PNGase F, peptide:*N*-glycosidase F; AP, alkaline phos; SPR, surface plasmon resonance; CSP, chemical shift perturbation.

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

0.9762 Å using an ADSC Q315R detector at ID29 at the ESRF. The data were indexed and integrated using DENZO and scaled with SCALEPACK (29). The crystals belong to the space group $P6_1$ with cell axes of $a = b = 54.4$ Å, $c = 56.1$ Å and diffracted to a resolution of 1.71 Å with one molecule in the asymmetric unit. The structure was determined by molecular replacement with PHASER (30) using the IPSE NMR structure with all the loops trimmed as a search model. The structure was refined using Refmac (31) with isotropic *B*-factors and manual rebuilding of protein was done with COOT (32). For more details see Table 1. The final model has no outliers in the Ramachandran plot, and 98.92% of all residues are in the favored region. A final check of the stereochemical quality of the final model was carried out using the program MolProbity (33).

Computational Modeling of a IPSEΔNLS·IgE Fab complex structure was performed using the ClusPro web server (34). The lowest energy structure of the IPSEΔNLS NMR structure ensemble and an IgE Fc structure from the protein database (PDB accession code 2WQR) were used for the rigid body docking. Unstructured terminal residues were removed from the IPSEΔNLS structure. Attractive restraints were introduced for residues that were significantly broadened or experienced chemical shift perturbations in the NMR titration. Repulsive restraints were used for unaffected residues.

Size Exclusion Chromatography—IPSE/α-1 dimer was mixed with IgE or IgG in a 1:1 molar ratio, preincubated for 30 min at room temperature, and applied to a Superdex 200 10/300 GL column (GE Healthcare). Chromatography was performed in PBS, pH 7.4, at 0.4 ml/min at the ÄKTA-Purifier 10 (GE Healthcare). Fractions of 0.5 ml were collected and analyzed by SDS-PAGE and silver staining.

Surface Plasmon Resonance—Experiments were performed on a BIAcore 3000 (GE Healthcare) and repeated at least twice. IgE or IgG was immobilized on a CM5 chip by standard amine coupling chemistry (for IPSE/α-1 dimer: IgE, 969 RU; IgG, 4435 RU; for IPSEΔNLS monomer: IgE, 5058 RU; IgG, 8428 RU). One flow cell was activated and deactivated without ligand for control purposes. Unspecific binding to the control flow cell was subtracted from the binding to the ligand-loaded flow cells. Running buffer was PBS, pH 7.5. Regeneration was performed with 0.075% SDS in PBS, pH 7.5. Measurements were performed under the following conditions: immobilized IgE, 100–700 nM IPSE/α-1 dimer with a flow rate of 10 μl/min for 10 min at 25 °C or 62.5–8000 nM IPSEΔNLS monomer with a flow rate of 10 μl/min for 10 min at 25 °C; immobilized IgG, 150–700 nM IPSE/α-1 dimer with a flow rate of 5 μl/min for 20 min at 25 °C or 125–16,000 nM IPSEΔNLS monomer with a flow rate of 10 μl/min for 10 min at 25 °C.

SDS-PAGE and Western blotting were performed according to standard protocols as described previously (7). Materials were obtained from the following suppliers: human IgG (Octagam); human IgE and human IgG Fab and Fc (Acris); human IgA (Dianova); human IgM (Medac). Immunoglobulins and immunoglobulin fragments were detected by alkaline phosphatase (AP)-labeled antibodies (diluted between 1:10,000 and 1:40,000) as follows: goat anti-human IgE, goat anti-human IgG (Fc), goat anti-human IgA, goat anti-human IgM, and goat anti-human Ig (κ light chain) (Dianova). Binding of IPSE/α-1 was detected by a monoclonal anti-IPSE/α-1 antibody 74–4B-7 followed by AP-goat anti-mouse IgG (Dianova) or using biotinylated IPSE/α-1 and subsequent incubation with AP-labeled streptavidin (SAV-AP) (Dianova). Dot blotting was performed using the SCR 96 D Minifold II-Dot Blotter (Schleicher & Schuell).

Sandwich dot blot was performed by dotting different amounts of human IgE (Acris) (4–0.5 pmol) onto nitrocellulose membrane. Following blocking of free binding sites, the membrane was incubated overnight either with IPSE/α-1 (125 pmol/ml), anti-IgE (30 pmol/ml; positive control), or with buffer (negative control). Following a washing step, membranes were incubated with biotinylated IgE (500 ng/ml) for 2 h and, subsequently, with alkaline phosphatase-labeled streptavidin diluted 1:10,000 for 1 h.

For biotinylation, 6 mg of IPSE/α-1 in PBS, pH 8.5, was mixed with 1 ml of a freshly prepared solution of *N*-hydroxysuccinimide-biotin (Sigma) in DMSO (1 mg/ml) and incubated for 4 h at room temperature followed by dialysis against PBS, pH 7.4. Biotinylated IPSE/α-1 was stored at −80 °C.

Peptide Scan—Overlapping 15-mer peptides were synthesized on cellulose membrane supports via the SPOT technique according to Roeckendorf *et al.* (35). After side chain deprotection of the peptides, the membrane was incubated for 10 min with 15 ml of 100% (v/v) ethanol and subsequently washed three times, for 10 min each time, with 15 ml of phosphate-buffered saline, pH 7.0 (PBS: 2.7 mM KCl, 1.5 mM KH_2PO_4, 136 mM NaCl, 8.1 mM Na_2HPO_4). For fake stripping, the membrane was incubated twice in an ultrasonic device, for 5 min each time, in 15 ml of *N,N′*-dimethylformamide, then for 5 min in 100 ml of double-distilled water, then twice, for 7 min each time, in 15 ml of denaturing buffer (8 M urea, 1% (w/v) SDS, 0.5% (v/v) β-mercaptoethanol·HOAc, pH 7.0), and finally three times, for 5 min each time, in PBS. The membrane was then blocked for 4 h with 25 ml of 1% (w/v) casein (Hammarsten grade, BDH) in PBST (PBS containing 0.05% (v/v) Tween 20). After a wash for 10 min with 20 ml of PBS, the membrane was incubated overnight in tight-fitting trays at 4 °C with 14.5 ml (2 ml/cm^2 membrane) of 6.7–8.3 nmol/ml IgE (Acris) that had been labeled according the manufacturer's instructions with DY781-*N*-hydroxysuccinimidyl ester (IgE-DY781; Dyomics) in blocking buffer. Subsequently, the membrane was washed five times, for 10 min each time, with 25 ml of PBST and then for 10 min with 25 ml of PBS. Fluorescence was quantitated and normalized to the mean fluorescence (=100%) of the parental peptide IPSE62–76 (Fig. 7*D*, *inset, spots D1* and *D2*) (Odyssey Infrared Imager, Odyssey software (version 1.2); Li-COR Biosciences).

Deglycosylation of Human IgG Fc Fragment by PNGase F Treatment—IgG Fc (Acris) was dialyzed against PBS, pH 7.5, containing 0.05 mM EDTA. 330 pmol of IgG Fc were incubated with 5000 milliunits of PNGase F (Roche Applied Science) in 100 mM sodium phosphate buffer containing 25 mM EDTA, pH 8.0, for 17 h at 37 °C. For control purposes, IgG Fc was incubated with buffer alone. SDS-PAGE and Western blotting were performed as described above. Carbohydrates were detected by AP-labeled *Aleuria aurantia* agglutinin (0.1 μg/ml; Vector Laboratories).

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

Double Immunodiffusion Assay (Ouchterlony)—The immunodiffusion assay was performed according to standard procedures. 1% agarose (HEEO ultra quality, Roth) in barbitone buffer, pH 8.2 (0.04 M 5,5-diethylbarbituric acid sodium salt and 0.017 M 5,5-diethylbarbituric acid, both Merck, conserved with 0.015% (w/v) merthiolate, Sigma), was pipetted onto carrier film (agarose gel support film; FMC). Using a gel punching device (CAMAG), holes were punched into the gel and each was filled with 8 μl of solution of binding partners (positions and concentrations see legend of Fig. 10*D*). After 48 h of incubation, the gels were pressed between two glass plates and several layers of pre-wetted filter paper with weights of 1 and 5 kg, respectively, for 10 min each. Gels were washed with 0.9% NaCl and distilled water, and dried at 40 °C before staining with Coomassie.

Induction of IL-4 Release from Human Basophils—Basophils were purified from peripheral blood of healthy human donors to a mean purity of 99% (36). Purified basophils were cultured and stimulated as described previously (7). Incubation was performed in 100 μl of culture medium (Iscove's modified Dulbecco's medium supplemented with 100 units/ml penicillin G, 50 μg/ml transferrin, 5 μg/ml insulin, 100 μg/ml streptomycin, and 10% fetal calf serum (PAA)) at a concentration of 10^6 basophils/ml. Cells were stimulated in the presence of IL-3 (2.5 ng/ml) as a basophil survival factor with various stimuli at the concentrations indicated, and supernatants were collected after 18 h and stored at −20 °C until further investigation. Interleukin-4 release was measured in the culture supernatants by using a two-site sandwich ELISA (Eli-Pair, Diaclone) according to the manufacturer's protocol.

Accession Codes—The atomic coordinates for the NMR ensembles and the crystal structures of IPSEΔNLS have been deposited in the Protein Data Bank under accession numbers 4AKA and 4EL6, respectively. The chemical shift assignments have been deposited in the Biological Magnetic Resonance Data Bank under accession number 17405.

Results

Solution and Crystal Structure of IPSE/α-1—IPSE/α-1 has recently been identified to be the major excretory/secretory component of *S. mansoni* eggs (37). However, apart from a translated cDNA sequence from *S. japonicum* (accession number FN317557) and recently sequenced genes from *S. hematobium* predicted to encode excretory/secretory proteins (11), a BLAST search in the sequence database did not reveal any similar sequences. On the primary structure level, a (Y/F)*XXXX*(Y/F)*XG* motif followed by a serine ~30 residues downstream suggested the presence of a greek key motif. Premature full-length IPSE/α-1 consists of 134 amino acids including an N-terminal signal peptide (residues 1–20), which is cleaved upon release of the mature IPSE/α-1 (residues 21–134). In the following, the residue numbering refers to the position in premature full-length IPSE/α-1. For this study, we cloned and expressed mature IPSE/α-1 (residues 21–134) and a monomeric deletion mutant of IPSE/α-1, IPSEΔNLS (residues 21–124), which lacks 10 amino acids at the C terminus, including a positively charged NLS (residues 125–131) (15) and Cys-132 being responsible for dimer formation (Fig. 1*A*). Expression in *Escherichia coli* of both wild-type IPSE/α-1 dimers and IPSEΔNLS monomers resulted in formation of inclusion bodies, which were solubilized by 8 M urea and refolded with physiological buffers containing the glutathione redox system GSH-GSSG to achieve dimerization. Two *N*-glycosylation sites present on the natural IPSE/α-1 were found not to be involved in basophil activation. *E. coli*-expressed nonglycosylated IPSE/α-1 exhibited the same functional activity as the glycosylated natural protein. Although neither the dimer nor the monomer expressed in *E. coli* were glycosylated, the wild-type IPSE/α-1 dimers aggregated after refolding at concentrations above 0.2 mg/ml. In contrast, the IPSEΔNLS monomers remained highly soluble at concentrations of >10 mg/ml and were thus suitable for structural studies (Fig. 1*B*).

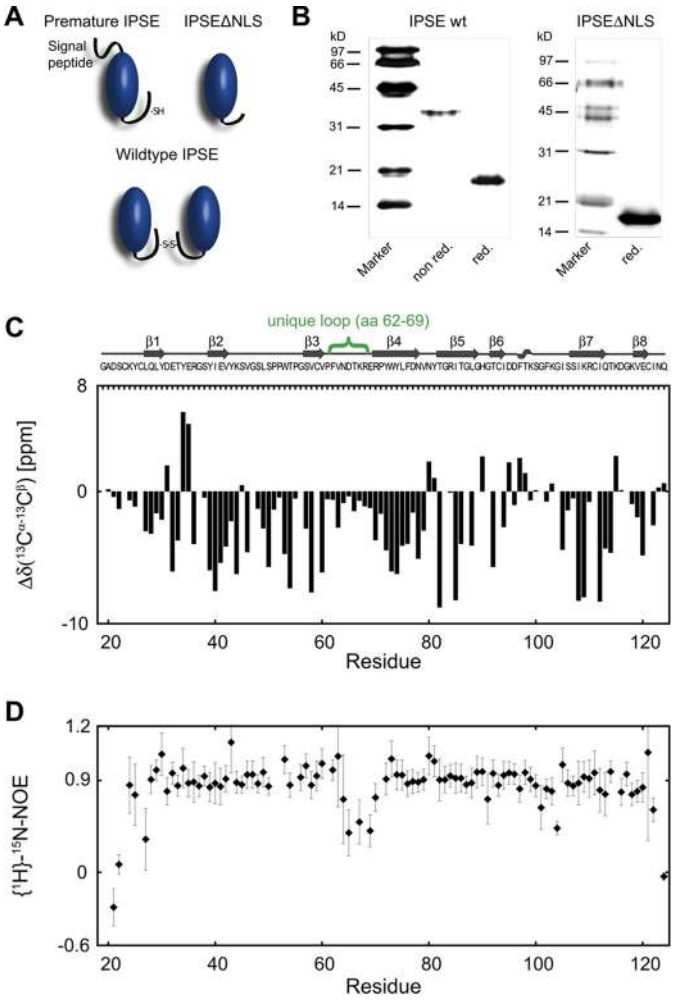


FIGURE 1. **Biochemical and NMR analysis of proteins studied.** *A*, schematic depiction of different mutants of IPSE/α-1. The monomeric IPSEΔNLS was used for structural analysis. *B*, comparison of recombinantly expressed and purified IPSE WT and IPSEΔNLS by SDS-PAGE and silver staining under nonreducing (*non-red*) and reducing (*red*) conditions. *C*, primary sequence, secondary structure, NMR ^{13}C secondary chemical shifts; *D*, $\{^1H\}$-^{15}N heteronuclear NOE of IPSEΔNLS.

Both the solution NMR structure and the crystal structure of IPSEΔNLS (Fig. 2*A*) show that the protein adopts a β-sandwich fold with two four-stranded β-sheets. The first sheet is formed by β-strands 1, 2, 3, and 8, and the second sheet is formed by β-strands 4, 5, 6, and 7. The two sheets are connected by a long loop between β-strands 3 and 4 (residues 62–69) and a short loop between β-strands 6 and 7 (residues 101–103). The two sheets are packed against each other in a hydrophobic core. The fold is stabilized by a disulfide bond between Cys-132 and Cys-24 (Fig. 2*B*). The crystal structure of IPSEΔNLS was solved by molecular replacement using the NMR structure as a search model. The crystal structure was refined to 1.8 Å resolution with an R_{work} of 0.19 and an R_{free} of 0.23 (Table 1). The crystal structure contains two molecules in the asymmetric unit, which are identical within the experimental error (root mean square deviation of 0.3 Å for 102 Cα atoms). The NMR structure ensemble is well defined with a root mean square deviation of 0.5 Å for backbone atoms and 1.0 Å for all heavy atoms (Table 1). The NMR structure ensemble and the crystal structure are identical within the experimental error (root mean square deviation of 1.0 Å for 102 Cα atoms).

To elucidate the molecular basis for the function of IPSE/α-1, we performed structural analysis of IPSEΔNLS using NMR spectroscopy and x-ray crystallography. NMR secondary chemical shift analysis indicated an all-β-strand secondary structure of IPSEΔNLS (Fig. 1*C*). ^{15}N NMR relaxation data

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

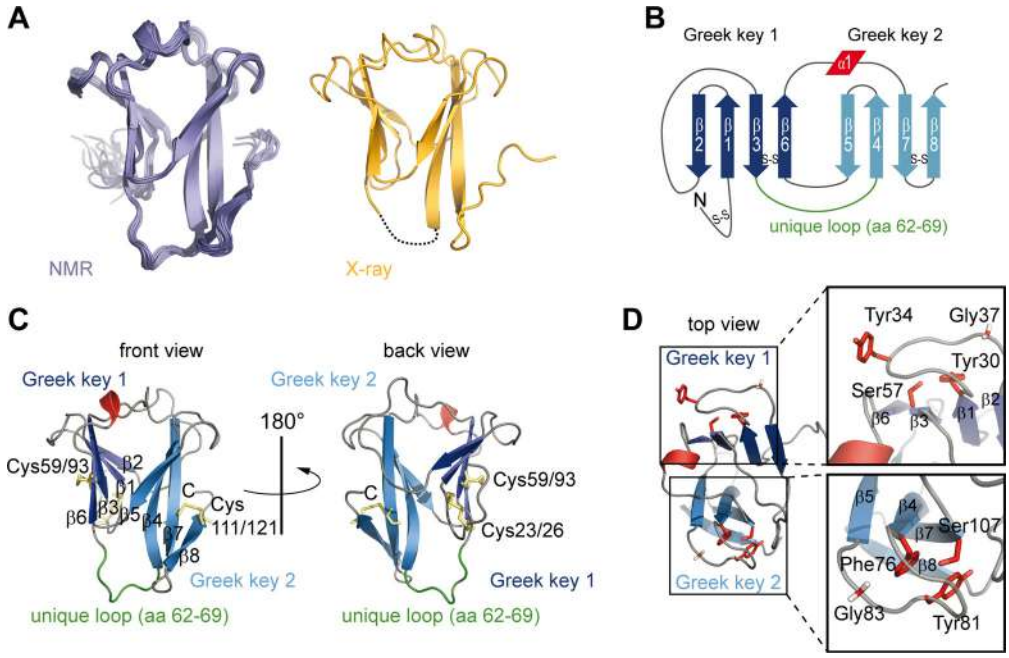


FIGURE 2. **Structure of IPSEΔNLS.** *A,* NMR (*left*) and crystal (*right*) structure of IPSEΔNLS. *B,* secondary structure topology of IPSEΔNLS. *C, ribbon* representation of the NMR-derived structure of IPSEΔNLS. Cysteine residues involved in disulfide bridges and the unique loop of the IPSE crystallin fold are indicated. *D,* zoomed view highlighting amino acids of the greek key signature sequence (*red sticks*).

showed that the polypeptide chain is rigid except for one loop (residues 62–69) that exhibits increased conformational flexibility (Fig. 1*D*). The tumbling correlation time $\tau_c \approx 7$ ns estimated from ^{15}N R_2/R_1 relaxation rates indicates that IPSEΔNLS (12 kDa) is a monomer in solution. This is in line with earlier studies demonstrating that the C-terminal cysteine (Cys-132) absent in IPSEΔNLS is required for covalent dimerization of wild-type IPSE/α-1 (16). The $^{13}C^{\beta}$ secondary chemical shifts of six other cysteines in IPSEΔNLS reflect an oxidized state of the attached sulfur atoms, consistent with the presence of three intramolecular disulfide bridges.

The solution structure of IPSEΔNLS (Fig. 2*A*, *left*) was determined based on a large number of NOE-derived distance restraints and dihedral angle restraints derived from TALOS+ (Table 1) (23). NOE patterns and chemical shifts indicated the presence of disulfide bonds between cysteines 23 and 26, 59 and 93, as well as 111 and 121, which served as additional restraints in the structural calculation.

Crystals of IPSEΔNLS, which diffracted to a resolution of 1.71 Å, were obtained as described earlier (17). The crystallographic structure (Fig. 2*A*, *right*, and Table 1) exhibited unambiguous electron density for the three disulfide bridges indicated above, whereas residues 64–68 within the flexible loop (see above) were not visible in the crystal structure due to missing electron density. The crystal and the NMR structure of IPSEΔNLS were in excellent agreement and superimposed with a backbone coordinate root mean square deviation (r.m.s.d.) of 0.5 Å. The structure of IPSEΔNLS consists of a sandwich of two β-sheets comprising eight anti-parallel β-strands (β1 to β8), which form two greek key motifs arranged in a pseudo 2-fold symmetry (Fig. 2*B*). The loops between the first two β-strands of each motif (β1/β2 and β4/β5, respectively) represent a characteristic β-hairpin motif. The two β-sheets are formed by three strands of one greek key motif each and a fourth β-strand contributed from the other greek key, as is typically observed in crystallin folds (Fig. 2*D*).

IPSE/α-1 Adopts a Crystallin Fold—The βγ-crystallin superfamily includes structural proteins from the vertebrate eye lens and a few distant relatives found in lower eukaryotes and prokaryotes. Most of the eye lens proteins contain four greek key motifs arranged in two domains. The structures of the C-terminal domain of human eye lens γ-B-crystallin (PDB code 2JDF), which is the best characterized member of the βγ-crystallin superfamily (Fig. 3*A*, *right*) and IPSEΔNLS monomer (Fig. 3*A*, *left*), are superimposed with a backbone coordinate r.m.s.d. of 1.7 Å. Despite this strong structural similarity at the primary sequence level, the conservation of IPSE and the C- or N-terminal domain of human eye lens γ-B-crystallin was less than 15% identity (Fig. 3*C*). Structural similarity despite low sequence identity has also been reported for several other members of the βγ-crystallin superfamily in lower organisms (38). Another example, Spherulin 3A from *Physarum polycephalum*, which contains only one domain with two greek key motifs, dimerizes via a disulfide bridge near the N terminus to restore the two-domain crystallin fold. IPSE/α-1 was found to dimerize in a similar fashion, although the respective cysteines are located at the C terminus rather than at the N terminus (Fig. 3*B*). Irrespective of the overall low sequence similarity, the highly conserved signature sequence (Y/F)*XXXX*(Y/F)*XG* represents a hallmark of the crystallin fold. The aromatic residues in this motif interact with a conserved serine on the third β-strand of each greek key motif (corresponding to β3 and to β7 of greek key motifs 1 and 2, respectively). Interestingly, in greek key motif 1 of IPSE/α-1, the spacing between the two tyrosine residues in the signature YDETYERG is shortened by one residue compared with other crystallin sequences (Fig. 3*C*).

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

TABLE 1
Refinement statistics for IPSEΔNLS

A) NMR Solution structure

	IPSEΔNLS
Distance constraints	
Total NOE	2650
Intra-residue	651
Inter-residue	
Sequential ($\|i-j\|=1$)	499
Medium-range ($\|i-j\|<4$)	286
Long-range ($\|i-j\|>5$)	1214
Intermolecular	0
Hydrogen bonds	0
Total dihedral angle restraints	161
ϕ	80
ψ	81
Structure statistics	
Violations (mean and s.d.)	
Distance constraints (Å)	0.009 ± 0.001
Dihedral angle constraints (°)	0.908 ± 0.053
Max. dihedral angle violation (°)	6.05
Max. distance constraint violation (Å)	0.18
Deviations from idealized geometry	
Bond lengths (Å)	0.016
Bond angles (°)	1.307
Impropers (°)	1.248
Average pairwise r.m.s. deviation** (Å)	
Heavy	0.77 ± 0.06
Backbone	0.32 ± 0.05

Pairwise r.m.s. deviation was calculated among 20 refined structures for residues 25-121

B) Crystal structure

	IPSEΔNLS
Refinement	
Resolution (Å)	47.1-1.71
No. reflections	10323
R_{work} / R_{free}	15.4 / 17.7
No. atoms	
Protein	798
Water	122
B-factors	
Protein	17.1
Water	27.3
R.m.s. deviations	
Bond lengths (Å)	0.011
Bond angles (°)	1.36

C) X-ray data-collection and processing statistics

X-ray source	ID29, ESRF
Wavelength (Å)	0.9762
Space group	P61
Unit-cell parameters (Å,°)	a = b = 54.40, c = 56.71, γ= 120
Observed reflections	83352
Resolution	47.11 - 1.71 (1.768 - 1.71)
Completeness %)	99.03 (92.07)
I/σ(I)	22.8 (3.66)
Multiplicity	7.4 (4.0)

As a result, the tyrosine rings cannot stack with each other so that Tyr-34 is solvent-exposed. However, the formation of the hairpin loop is not significantly affected (Fig. 2*D*).

Although the fold of the human lens γ-B-crystallin is maintained by strong hydrophobic interactions in the protein core, the IPSE fold is stabilized by three disulfide bonds (23/26, 59/93, and 111/121) (Fig. 2, *B* and *C*), consistent with its high stability over a wide pH range and against 2 M urea. These disulfide bonds may compensate for the lack of a tyrosine corner (39), which is present in most γ-crystallins (see γ-B-crystallin, Tyr-151 in Fig. 3*A*, *right*). A typical feature of a tyrosine corner is a hydrogen bond between the hydroxyl group of the tyrosine side chain (Tyr-151 in Fig. 3*A*, *right*) and the peptide backbone two to five amino acids upstream stabilizing the formation of a β-hairpin. In contrast, the IPSE/α-1 fold and stability appear to depend on disulfide bridge formation, consistent with the observation that IPSEΔNLS rapidly precipitated under reducing conditions. Distinct from other proteins of the crystallin family, the β7 and the β4 strands of greek key motif 2 in IPSE/α-1 are elongated (Fig. 3*C*). Furthermore, orthogonal stacking of Tyr-74 and Tyr-72 and additional hydrophobic contacts lead to a curvature of this β-sheet (Fig. 3*A*, *left*). The long flexible loop between the two greek key motifs (residues 62–69; Fig. 3, *A* and *C*, colored in *green*) turned out to be a unique feature of IPSE/α-1 compared with other crystallins.

IPSE/α-1 Is an Immunoglobulin-binding Factor—Earlier, we demonstrated that the presence of IgE on the surface of basophils is essential for functional activity of IPSE/α-1. This prompted us to investigate the interaction of IPSE/α-1 with IgE, other immunoglobulin isotypes, and further serum components. All experiments described in the following were performed with the IPSE/α-1 dimer, which is the basophil-activating form. Western blots from whole human serum samples demonstrated that IPSE/α-1 binds to the immunoglobulin bands and to a band identified by the N terminus sequencing as transferrin, but not to other serum components (Fig. 4*A*). More detailed analysis confirmed that it binds to all human serum immunoglobulin isotypes (IgE, IgG, IgA, and IgM) in a nonantigen-specific manner (Fig. 4*B*). Thus, IPSE/α-1 is a general immunoglobulin-binding factor. Titration of IgE and IgG on dot blots suggested that IPSE/α-1 binds with higher affinity to IgE than to IgG (Fig. 4*C*). This finding was corroborated by size exclusion chromatography of preformed IPSE/α-1·IgE and IPSE/α-1·IgG complexes in solution. IPSE/α-1·IgG, in contrast to IPSE·α-1-IgE, completely dissociated upon the mechanical shear forces during size exclusion chromatography. This demonstrated that IPSE·α-1 binds with considerably lower affinity to IgG than to IgE (Fig. 5).

To quantify the binding affinity of IPSE/α-1 to IgE and IgG, we performed surface plasmon resonance (SPR) assays with IPSE/α-1 dimer and IPSEΔNLS monomer. SPR assay with the IPSE/α-1 dimer (Fig. 6*A*) showed that IPSE/α-1 clearly bind to immobilized immunoglobulins on the CM-5 sensor chip. However, the binding isotherms obtained upon analysis of the data showed a sigmoidal curve indicating a complex binding scenario of dimeric IPSE/α-1 to the dimeric immunoglobulins. As a consequence, simple curve fitting using the Langmuir (1:1) calculation was not applicable to determine quantitative affinity data. Thus, we performed additional SPR experiments with the monomeric mutant IPSEΔNLS (Fig. 6*B*). Fitting of the binding isotherms using a 1:1 binding model revealed a K_D value of 1.7 μM for the binding of IPSEΔNLS to IgE and 7.4 μM for the binding to IgG, which corresponds to a 4-fold higher binding affinity to IgE compared with IgG. This corroborates the earlier observations that IPSE/α-1 binds with higher affinity to IgE than to IgG, although one should keep in mind that the overall binding affinity of the dimer to IgE might differ from that of the monomeric mutant.

Analysis of the Interaction Sites of IPSE/α-1 with IgE—To map the binding interface of IPSE/α-1 with the IgE antibody, we compared NMR chemical shifts of IPSEΔNLS free and IPSEΔNLS bound to IgE at molar ratios of 0.5:1 and 1:1

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

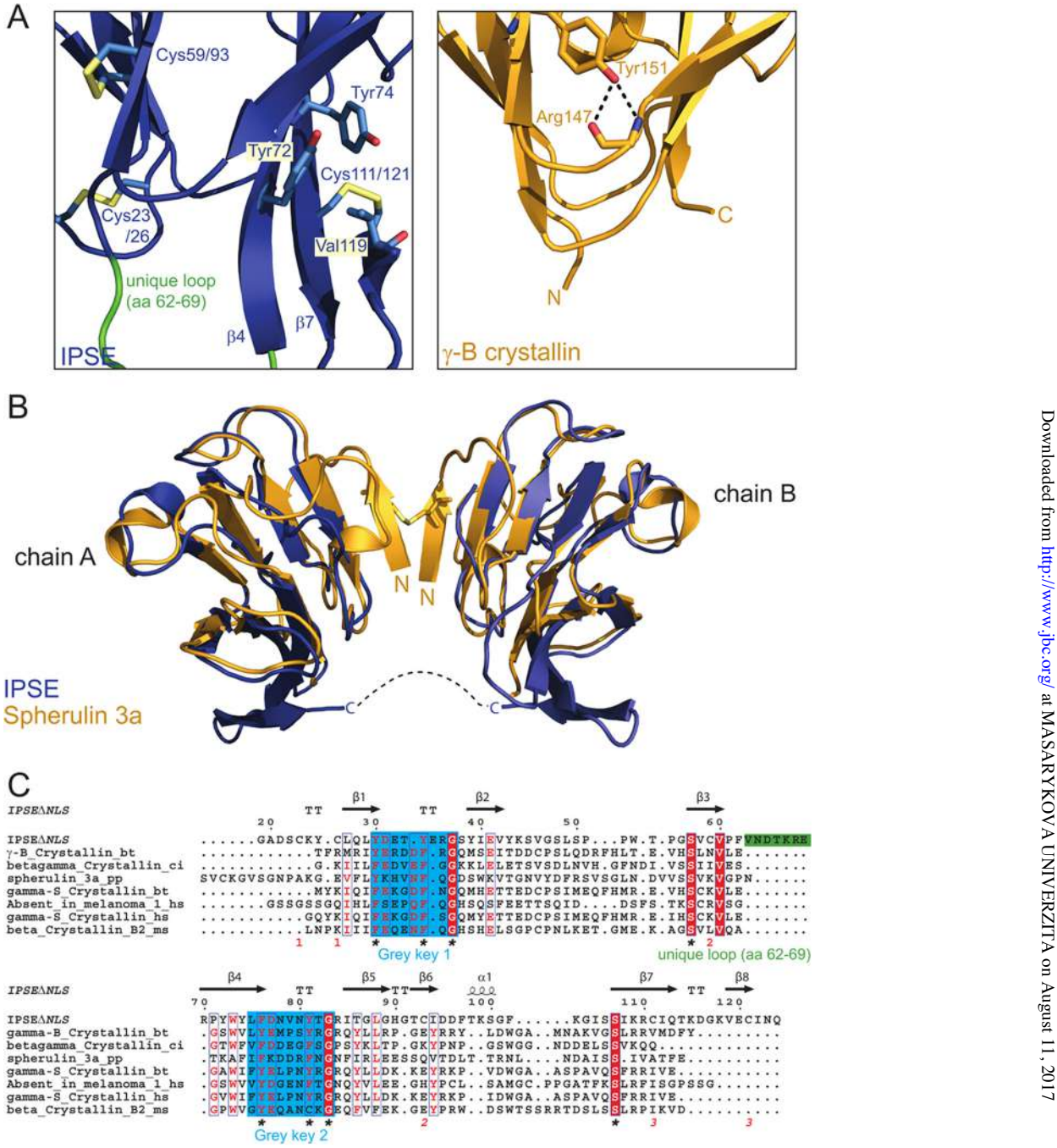


FIGURE 3. **IPSEΔNLS adopts a $\beta\gamma$-crystallin fold.** *A*, zoomed view of IPSEΔNLS (*blue*, *right*) and human eye lens γ-B-crystallin (*dark yellow*, *left*). Shown are the residues of the tyrosine corner of γ-B-crystallin (Tyr-151 and Arg-147) and the orthogonal stacked Tyr-72 and -74 as well as the disulfide bridges of IPSEΔNLS (*sticks*) and the flexible loop (amino acids (*aa*) 62–69) (*green*). *B*, structure superposition of two IPSEΔNLS monomers (*blue*) with dimeric Spherulin 3A (*orange*). Structures align with a backbone r.m.s.d. of 1.8 Å. *C*, structure-based multiple sequence alignment of crystallin superfamily members. *Above* the sequences: secondary structure elements of IPSEΔNLS (except of disulfide bridges). *Below* the sequences: disulfide bridges are shown as *red numbers*; the flexible loop (amino acids 62–69) unique to IPSE/α-1 is depicted as a *green bar*; the two (F/Y)*XXXX*(F/Y)G signature sequences and the conserved serine residues characteristic of the crystallin fold are indicated by *asterisks*.

(IPSEΔNLS/IgE). Given the high molecular mass of the complex (≈200 kDa), ^{2}H labeling of IPSEΔNLS was required for sufficient sensitivity. Uniformly $^{2}H/^{15}N/^{13}C$-labeled IPSEΔNLS was titrated with unlabeled human IgE to monitor chemical shift perturbations (CSP) using $^{1}H,^{15}N$ correlation experiments. Upon binding of IPSEΔNLS to the 180-kDa IgE molecule, IPSEΔNLS signals experience extensive line broadening, which likely results from the low molecular tumbling rate and thus strongly increased NMR line widths in the high molecular weight complex. We therefore used relaxation-optimized NMR experiments (28) to detect IPSEΔNLS bound to IgE (Fig. 7*A*). A number of amide proton signals close to the interaction site remain broadened beyond detection, although for several other signals significant chemical shift differences are observed compared with free IPSEΔNLS. The binding kinetics of the IPSE/IgE interaction is slow at the NMR chemical shift time scale

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

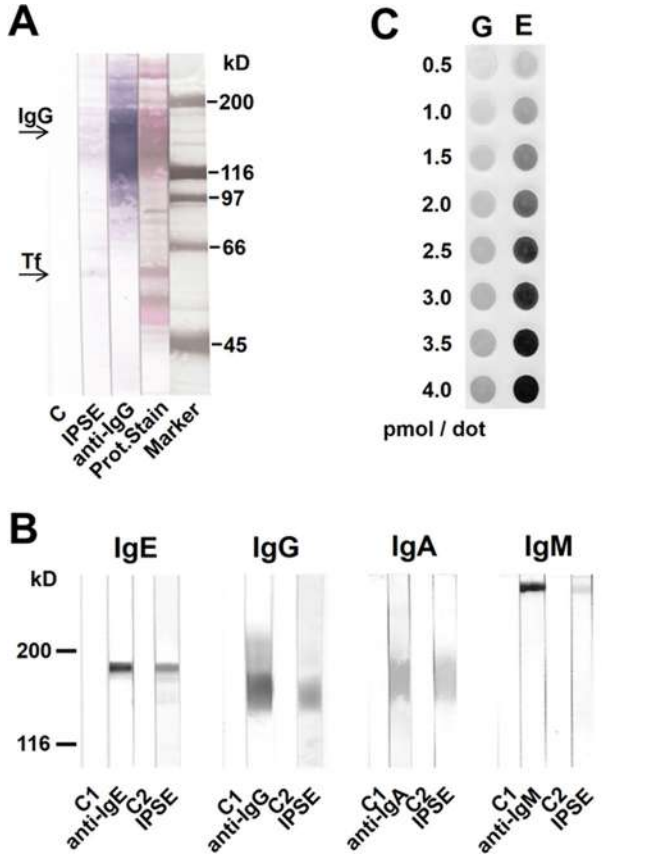


FIGURE 4. **IPSE/α-1 is an immunoglobulin-binding protein with highest affinity to IgE.** *A,* IPSE/α-1 binds to immunoglobulins and transferrin (*Tf*) in human serum. Western blots of human serum were incubated with IPSE/α-1 followed by labeled streptavidin (SAV-AP) or labeled α-human IgG. Proteins were stained with AuroDye and india ink. *B,* IPSE/α-1 binds to various immunoglobulin isotypes. Strips with blotted human polyclonal IgE, IgG, IgA, and IgM were incubated with the respective anti-isotype antibodies or IPSE/α-1 followed by SAV-AP. Controls: secondary antibody (C1), SAV-AP (C2). *C, dotted* IgG (*G*) and IgE (*E*) were incubated with labeled IPSE/α-1 followed by SAV-AP.

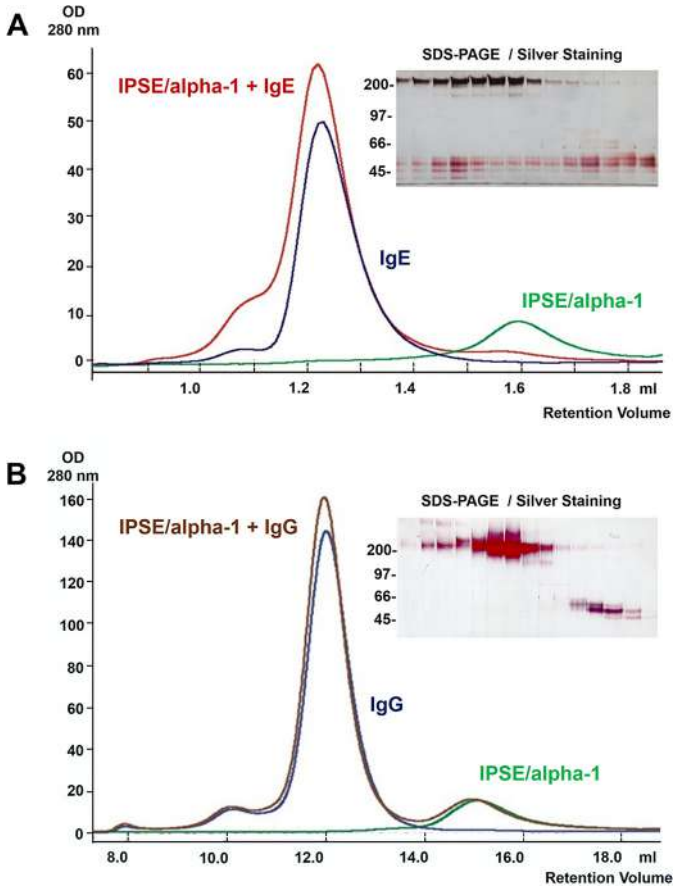


FIGURE 5. **Size exclusion chromatography.** Complexes of IPSE/α-1 and IgE (*A*) and complexes of IPSE/α-1 and IgG (ratio 1:1), respectively (*B*), were applied to a Superdex 75 column (*red curve*). For control, the same amount IPSE/α-1 alone (*green curve*) and IgE or IgG alone (*blue curve*) were run over the column. Fractions of the IPSE/α-1·IgE or IPSE/α-1·IgG complexes were analyzed by SDS-PAGE and silver staining. High molecular weight fractions contain IPSE/α-1 and IgE but not IgG, suggesting that IPSE/α-1 binds with higher affinity to IgE than to IgG.

consistent with a low micromolar to nanomolar dissociation constant. The observed CSPs identify a large binding surface on one side of IPSEΔNLS that includes the flexible loop (residues 62–69) connecting the two greek key motifs (Fig. 7, *A–C*). The opposite side of the structure (β1 and β2 with the hairpin loop in greek key motif 1, and β7 and β8 in greek key motif 2) is largely unaffected by IPSE binding to IgE. The amide group of Val-63, which is in close proximity to the solvent-exposed aromatic side chain of Phe-62 in the flexible loop, experiences the strongest chemical shift perturbation. Another patch of strongly affected residues, *e.g.* Gly-83, is in close spatial proximity to the highly conserved Tyr-81. The NMR data thus indicated that the binding interface involves these aromatic side chains. Most of the interface is located in overall positively charged regions of the protein, especially the flexible loop (residues 62–69) and β-strand β4. This implies an essential role of electrostatic and charged interactions.

Peptide scan analysis of the sequence requirements in the peptide comprising residues 62–76 revealed that exchanges of positively charged amino acids to either alanine or glutamate lead to a drastic decrease in IgE binding (B1, B2, B4 and D3, D4, D5, respectively) (Fig. 7*D*). This highlights the importance of the positive charges for the IPSE/IgE interaction. In addition, exchange of the aromatic residues Tyr-72 and -74, Trp-73, Phe-76, and Leu-75 by alanines (C1, C2, C3, C4, and C5) strongly decreases the binding affinity to IgE, demonstrating the importance of additional hydrophobic interactions. Interestingly, scrambling of the amino acid sequence did not change the binding intensity significantly, suggesting that the overall charge and hydrophobicity might be even more important than the sequence itself. Taken together, the peptide scan and NMR titration demonstrated the importance of positively charged and aromatic amino acid side chains for enabling effective binding of IPSE/α-1 to IgE.

The residues relevant for the interaction of IPSE/α-1 with IgE as determined by NMR and peptide scan experiments were further investigated by mutational analysis (Fig. 8). Mutations were introduced both into IPSEΔNLS (monomeric due to lack of the inter-chain disulfide bond) and into wild-type IPSE/α-1 (dimeric). The monomeric IPSEΔNLS mutants were analyzed by NMR for structural integrity (Fig. 9, *A* and *B*); the dimeric IPSE/α-1 mutants were analyzed for immunoglobulin binding (Western blotting; Fig. 8*A*) and for basophil IL-4-releasing capacity (Fig. 8*B*).

The mutants I108E/K109E, Q28R/E41K, and D31K/E32K showed reduced immunoglobulin binding and loss of func-

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

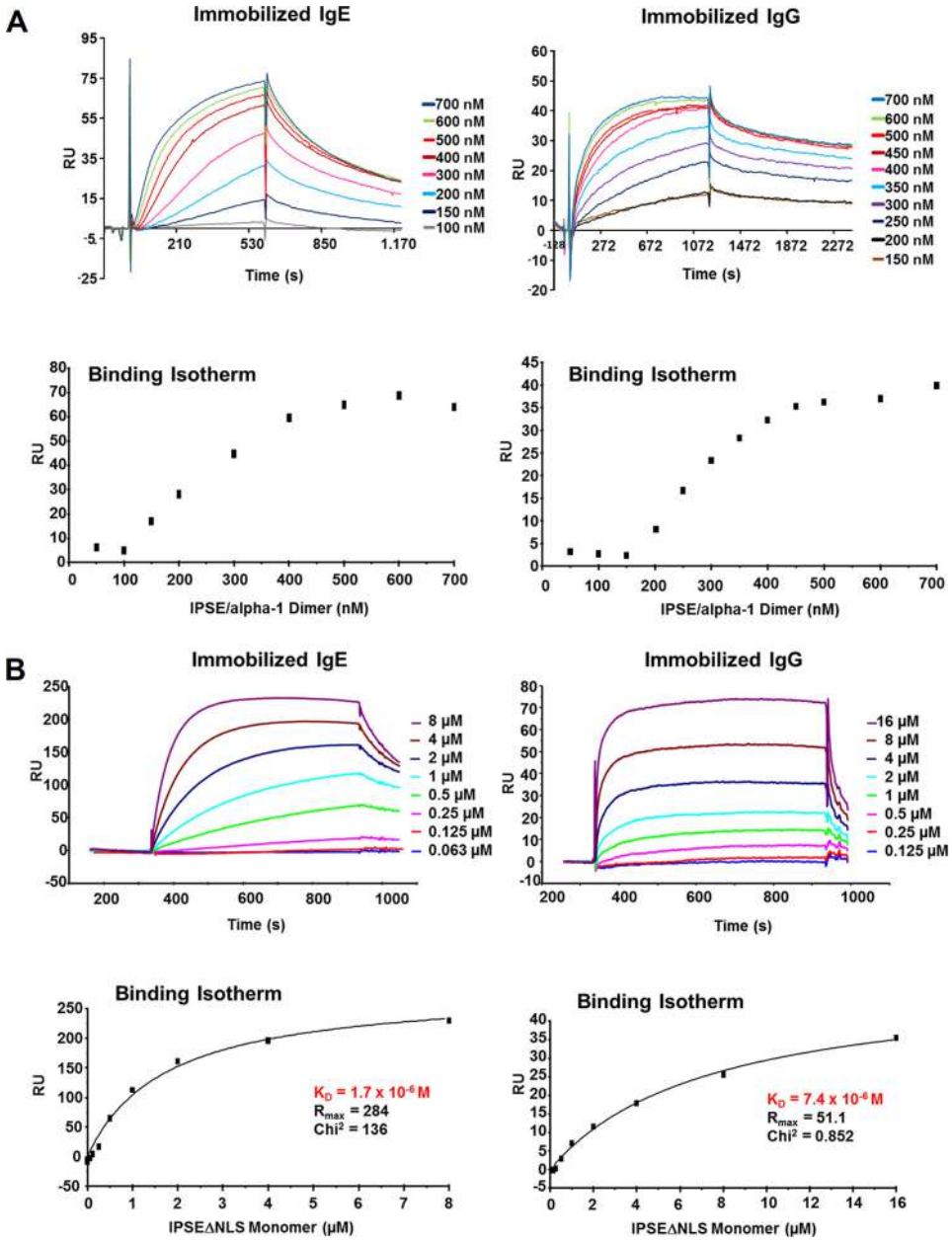


FIGURE 6. **Surface plasmon resonance assays.** Binding of IPSE/α-1 dimer (*A*) and IPSEΔNLS monomer (*B*) to IgE and IgG immobilized on a CM5 chip was analyzed by real time measurement. *Curves* were obtained after subtraction of the unspecific binding to the reference cell. Bound IPSE (*RU*) was estimated at the end of the association phase. Binding isotherms of the IPSE/α-1 dimer (*A*) showed a sigmoidal curve characteristic for multivalent positive cooperative binding. Thus, Langmuir (1:1) calculation of affinity constants was not applicable. For the IPSEΔNLS monomer (*B*) dissociation constants of 1.7 and 7.4 μM for IgE and IgG, respectively, were measured.

tional activity due to compromised *in vitro* folding. NMR analysis revealed that the monomeric IPSEΔNLS mutant I108E/K109E is unfolded in solution (Fig. 9*A*). In contrast, the monomeric mutants Q28R/E41K and D31K/E32K appeared folded in the NMR analysis. However, Western blot analysis of the corresponding dimeric mutant preparations revealed a shift in the dimer/monomer ratio toward the monomer (Fig. 8*A*). Thus, the reduced functional activity of these two preparations is potentially due to a reduced proportion of the functionally active dimer.

All other IPSE/α-1 mutants were structurally intact as assessed by NMR (monomeric IPSEΔNLS mutants) and by monitoring the intensity of dimeric IPSE/α-1 bands on Western blots. The mutants N80A/T82A and K67E/R68E displayed immunoglobulin binding and IL-4-inducing capacities similar to the wild-type IPSE/α-1. This is consistent with the fact that the respective residues are located outside (Asn-80 and Thr-82) or at the edge (Lys-67 and Arg-68) of the binding interface as assessed by NMR (Fig. 9*C*). The double mutant F62A/N64A exhibited reduced binding to IgG and IgE compared with the wild type and markedly reduced basophil-activating capacity. The pronounced effect of the amino acid exchange at positions 62 and 64 corroborates the critical role of the flexible loop (amino acids 62–69) as the core of the immunoglobulin interaction site. The mutant T92Y/R127L is identical with a factor from *S. mansoni* eggs termed SmCKBP (40). SmCKBP has been

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

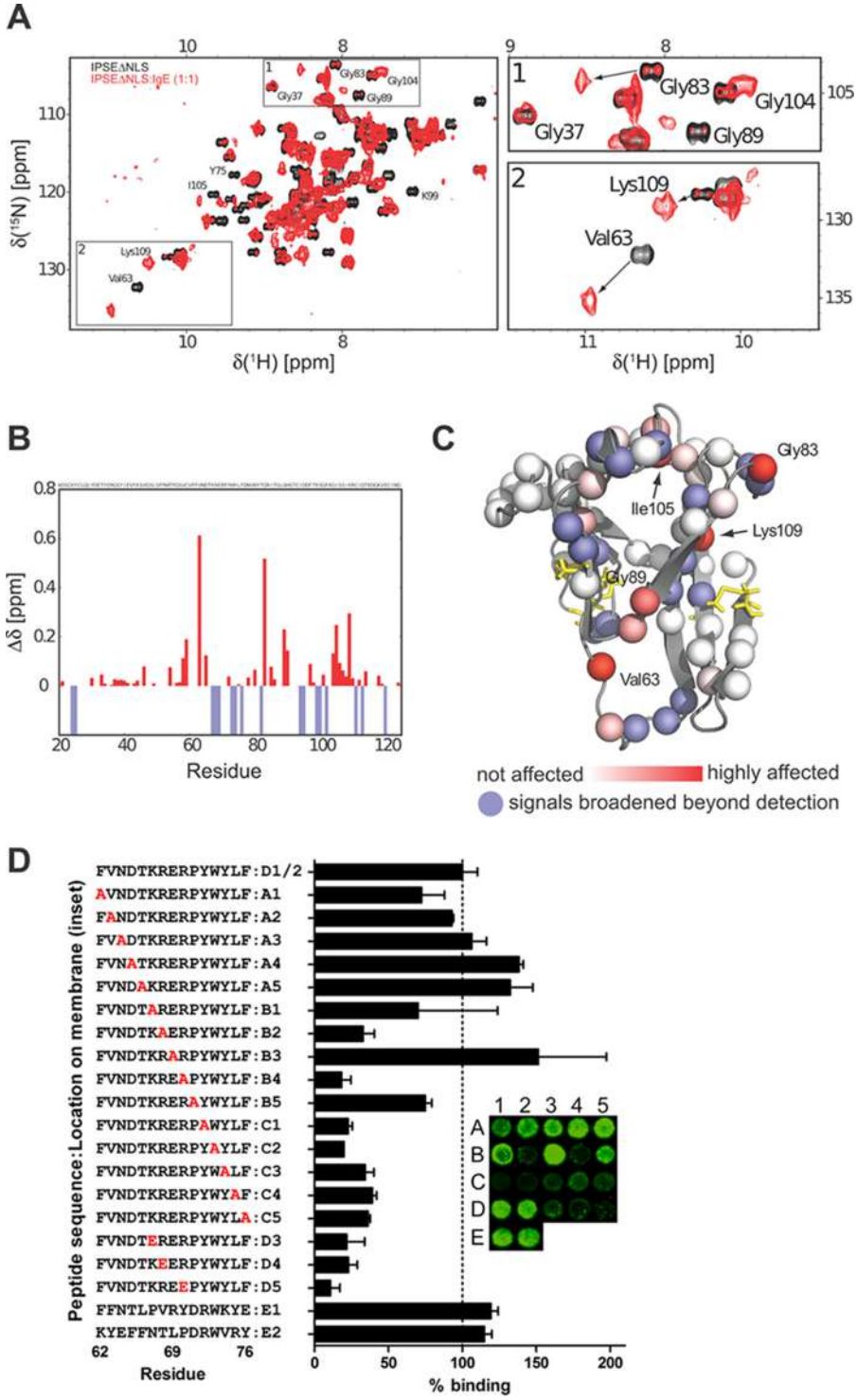


FIGURE 7. **Mapping the binding surface of IPSEΔNLS with IgE.** *A,* $^{1}H,^{15}N$ NMR correlation spectra of IPSEΔNLS (*black*) and the IPSEΔNLS·IgE complex (molecular mass ~200 kDa, *red*). The presence of NMR signals corresponding to free IPSEΔNLS at a molar ratio of ~1:1 (IPSEΔNLS/IgE) suggests that IgE has only one IPSEΔNLS-binding site. (Note, the NMR signals of even a very small fraction of free IPSEΔNLS (10 kDa) are much stronger than those of the 200-kDa complex.) *B,* NMR chemical shift perturbation (CSP) *versus* residue number. *Blue negative bars* indicate amide signals that are broadened beyond detection. *C,* mapping of the IgE binding interface onto the structure of IPSEΔNLS. Nitrogen atoms of the backbone amide groups are represented as *spheres* and colored *red* according to CSP or *blue* if line broadening is observed in the NMR titration experiment. *D,* peptide scan. Binding of labeled IgE was to immobilized 15-mer IPSE-derived peptides. Binding (means ± S.E.) of IgE-DY781 to 15-mer peptides is expressed in relation to the IPSE-derived peptide 62–76 (*inset, spots D1* and *D2*; mean fluorescence = 100%). Residues exchanged to Ala or Glu in the respective peptides are highlighted in *red*. *Inset,* representative example of a peptide library read-out.

reported to bind to, and neutralize, chemokines but to be unable to activate basophils. In accordance with this previous report, recombinant T92Y/R127L displayed markedly reduced immunoglobulin binding together with a complete lack of IL-4-releasing capacity. These results were consistent with the position of Thr-92 within the binding surface as mapped by the

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

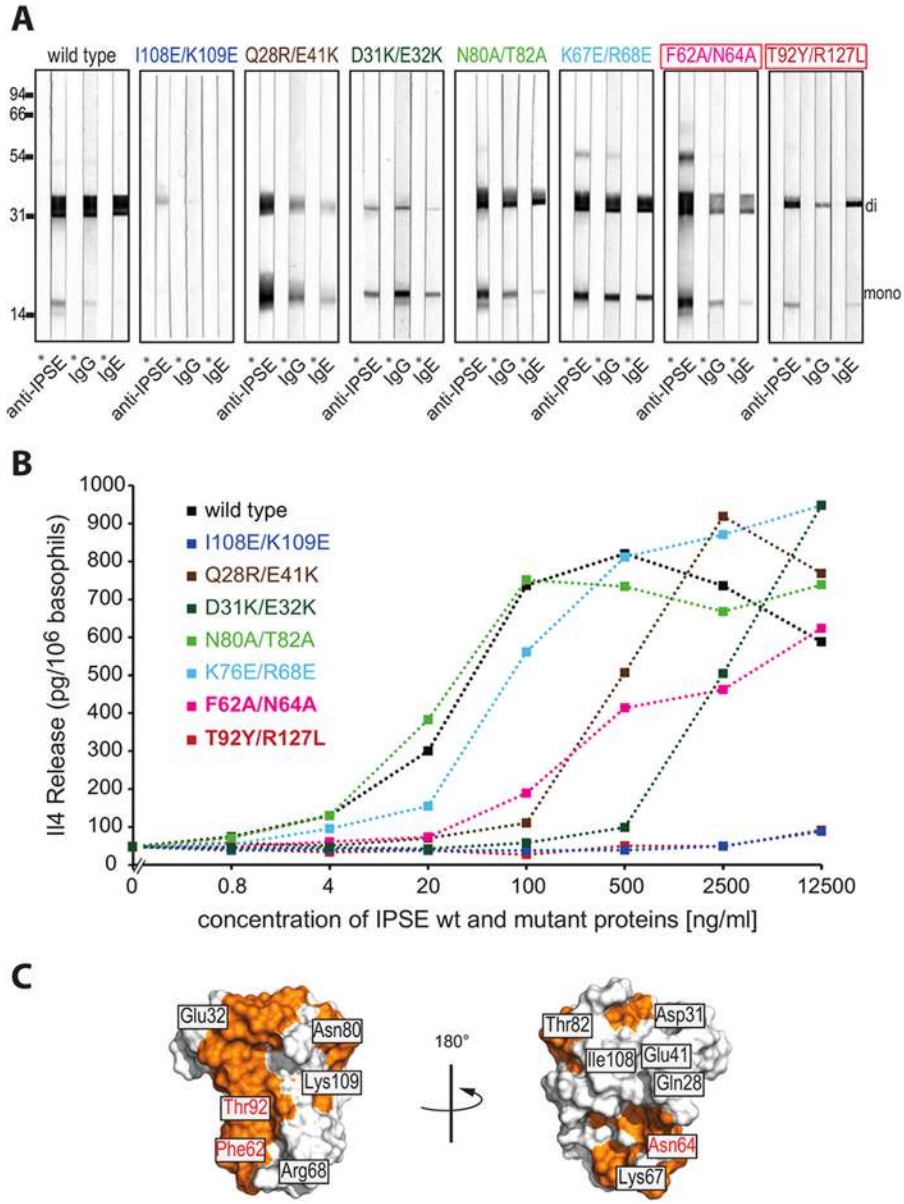


FIGURE 8. **Mutational analysis of immunoglobulin binding and functional activity of IPSE/α-1.** *A,* binding of monoclonal anti-IPSE antibody, human IgG, or human IgE to IPSE/α-1 wild-type and mutants, respectively, assessed by Western blots. Mutants in *red boxes* were analyzed by SPR for binding affinity with IgE. *B,* IL-4 release from human basophils induced by different concentrations of IPSE/α-1 wild-type and mutants, respectively. Shown is one representative result of three independent experiments. *C,* surface representation of IPSEΔNLS. Binding interface was derived from the NMR titration and the peptide scan (*orange*).

NMR titrations (Figs. 8 and 9*C*). Taken together, the mutational analysis showed a good correlation between the ability of IPSE/α-1 to bind immunoglobulins and to activate basophils. Our results suggest that IgE binding involves a binding surface that includes the unique flexible loop of the IPSE/α-1 crystallin fold, although functional activity requires the presence of dimeric IPSE/α-1.

IPSE/α-1 Does Not Cross-link IgE—The finding that only dimeric but not monomeric IPSE/α-1 is able to activate basophils suggests that the functional activity of this molecule involves cross-linking of receptor-bound IgE, the classical mechanism seen with allergens, lectins, and B cell superantigens. Because an antigen-specific binding mode (such as exhibited by an allergen) could be excluded, we asked whether IPSE/α-1 interacts with the carbohydrate side chains (like a lectin) or with the Fab or Fc regions of immunoglobulins (like a B cell superantigen). We therefore performed Western blot experiments with Fab and Fc fragments of human IgG. IgG was used as a paradigmatic immunoglobulin, as this isotype is available in large quantities and is bound by IPSE/α-1.

A lectin-like binding of IPSE/α-1 to immunoglobulin glycan side chains could be excluded by comparing its binding to native (*i.e.* glycosylated) and to PNGase F-deglycosylated Fc fragments of IgG (Fab fragments of IgG are not glycosylated). Deglycosylation of IgG Fc was confirmed by Western blot analysis, as deglycosylated IgG Fc was found to band at a lower molecular weight than native IgG Fc (Fig. 10*A*). Moreover, the lectin *A. aurantia* agglutinin that binds to fucose residues on

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

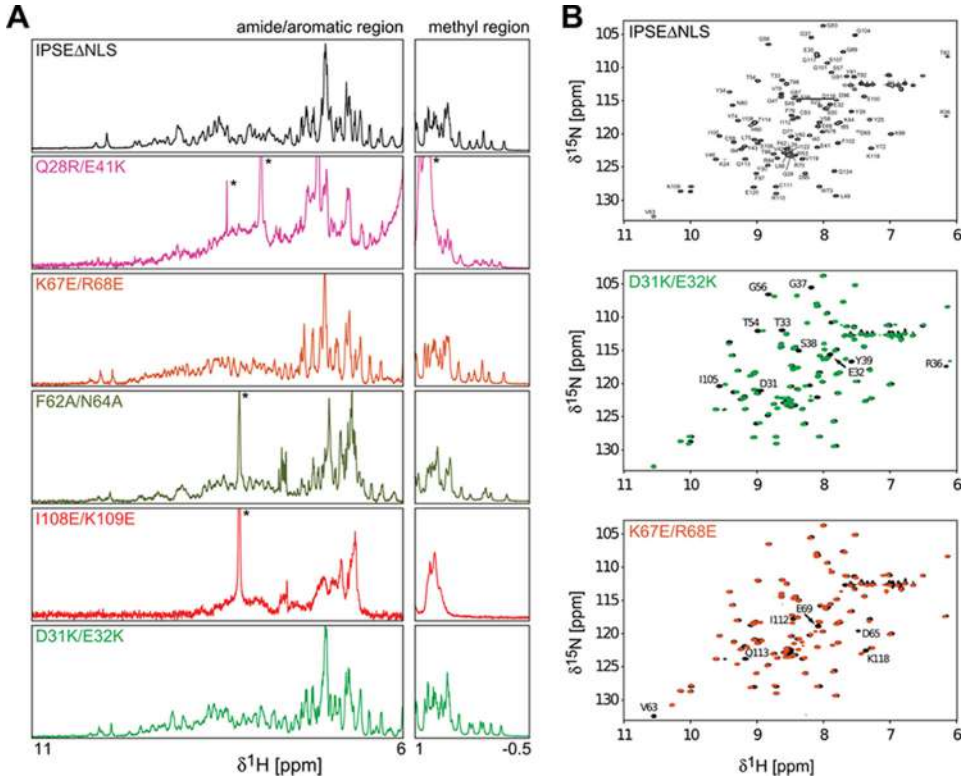


FIGURE 9. **One-dimensional ^{1}H NMR (*A*) and ^{1}H,^{15}N HSQC spectra (*B*) spectra of IPSEΔNLS and IPSEΔNLS mutants.** *A,* spectral regions of amide and aromatic and methyl protons for the mutants Q28R/E41K, K67E/R68E, F62A/N64A, I108E/K109E, and D31K/E32K are shown. The good dispersion of amide proton signals and the presence of methyl proton signals below 0 ppm observed for all variants except I108E/K109E are indicative of the presence of tertiary folds. Signals arising from low molecular weight contaminations are marked with an *asterisk*. Spectra were recorded at 600 MHz ^{1}H Larmor frequency. *B,* only signals of amides in close spatial proximity to the mutations are shifted in the spectra of D31K/E32K and K67E/R68E mutants. Spectra were recorded at 600 MHz ^{1}H Larmor frequency.

native IgG no longer detected deglycosylated IgG Fc. In contrast, IPSE/α-1 recognized both glycosylated and deglycosylated IgG Fc. This clearly rules out a lectin-like interaction between IPSE/α-1 and IgG, a finding that quite likely applies also to the other immunoglobulin isotypes (due to the highly conserved immunoglobulin domains).

To investigate whether IPSE/α-1 binds to the Fab or the Fc part of IgG, whole IgG and Fab and Fc fragments were analyzed for binding of IPSE/α-1 by Western blotting. IPSE/α-1 was found to bind to both Fab and Fc fragments (Fig. 10*B*).

Typically, IgE-dependent activation of basophils occurs via cross-linking of IgE on their surface as described for allergens, anti-IgE antibodies, lectins, or B cell superantigens. Thus, we expected the same mechanism for IPSE/α-1. However, methods to demonstrate cross-linking, such as sandwich dot blots and double immunodiffusion assay, showed that IPSE/α-1 was not able to cross-link IgE. By using sandwich dot blots, we investigated whether IPSE/α-1 cross-linked unlabeled membrane-bound IgE with labeled soluble IgE. Although the blot with anti-IgE (positive control) showed a clear positive signal, the blot with IPSE/α-1 signal did not exceed background (buffer control) levels indicating that IPSE/α-1 was not able to cross-link IgE (Fig. 10*C*). A background signal observed with ≥3.3 pmol pf IgE/dot likely resulted from the tendency of IgE to form aggregates at high concentrations. Analogous results were obtained with the double immunodiffusion assay. Binding partners were applied in different concentration ratios to wells punched in the agarose on microscope slides. After diffusion into the gel for 48 h, immunoprecipitation due to cross-linking of the binding partners was visualized by Coomassie staining. As expected, all ligand combinations used as positive controls (IgE/anti-IgE, IgG/anti-IgG, and IgG/protein L (a B cell superantigen from *Peptostreptococcus magnus*) and IPSE/α-1/anti-IPSE antiserum) yielded clearly visible precipitin lines, but neither IPSE/IgE nor IPSE/IgG formed complexes large enough to precipitate in the gel, suggesting that IPSE/α-1 did not cross-link IgE or IgG in solution (Fig. 10*D*). Of note, the small IPSE/α-1·IgE complexes formed at equivalence as assessed by size exclusion chromatography indicated a binding stoichiometry of one IPSE/α-1 dimer to one IgE molecule and thus confirmed that cross-linking did not take place (Fig. 5).

Discussion

Accumulating evidence suggests an important role of basophils for the immune response against helminth infection. However, the molecular mechanism of parasite-induced basophil activation is unknown for the majority of worm infections. In the case of schistosomiasis, IPSE/α-1 represents the active principle responsible for activation of basophils (7). To better understand the molecular mechanisms by which IPSE/α-1 induces basophil IL-4 release from nonsensitized healthy individuals, we determined the structure of this molecule and investigated its interaction with IgE.

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

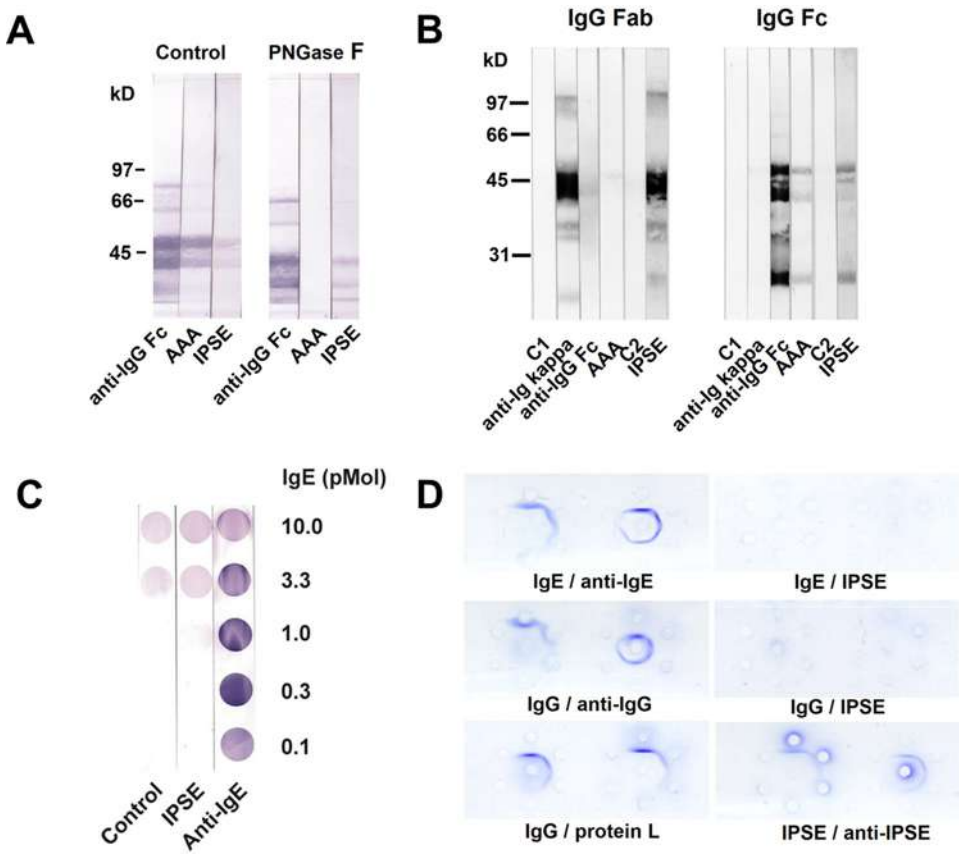


FIGURE 10. **Mode of interaction of IPSE/α-1 with immunoglobulins.** *A,* binding of IPSE/α-1 to deglycosylated IgG Fc. Blotted mock-treated IgG Fc (*Control*) and IgG Fc deglycosylated with PNGase F were incubated with α-human IgG Fc or *A. aurantia* agglutinin or labeled IPSE/α-1. *B,* IPSE/α-1 binds to Fab and Fc fragments. Blotted Fab and Fc fragments of human IgG were detected by α-human IgG (*Fc*), by α-κ light chain (*anti-Ig kappa*), by *A. aurantia* agglutinin or by labeled IPSE/α-1. *C1* = buffer control and *C2* = SAV-AP. *C,* sandwich blots. Human IgE dotted at different amounts (10 to 0.1 pmol) was incubated with buffer (negative control) or α-IgE (positive control) or IPSE/α-1, followed by biotinylated IgE. α-IgE gives a clear signal with a maximum at 0.3 pmol/dot, whereas IPSE/α-1 does not exceed background level. *D,* double immunodiffusion assay (Ouchterlony). Binding partners were applied at various ratios. One partner was pipetted into the center well, the other one clockwise beginning at 12 o'clock in order of dilution in the surrounding wells: 1:1, 1:2, 1:4, 1.8, 1:16, and 1:32. The following starting concentrations were used: 0.5 mg/ml IgE and 1 mg/ml anti-IgE; 0.5 mg/ml IgG and 0.5 mg/ml anti-IgG; 0.5 mg/ml IgG and 0.5 mg/ml protein L (a B cell superantigen from *P. magnus*); 0.25 mg/ml IgE and 0.05 mg/ml IPSE/α-1 (corresponds to a molar ratio of ~1:1); 0.25 mg/ml IgG and 0.05 mg/ml IPSE/α-1; and 0.05 mg/ml IPSE/α-1 and rabbit anti-IPSE antiserum (1:2). Precipitation *arcs* indicate complex formation.

The solution and crystal structures of the monomeric IPSEΔNLS classified IPSE/α-1 as a novel member of the βγ-crystallin superfamily. βγ-Crystallins, such as the eye lens γ-B-crystallins, are characterized by extremely high stability toward harsh conditions such as extreme pH (pH 1), high temperature (70 °C), and chaotropic agents (6 M urea) (38). For IPSE/α-1, such a stable structure might be important, because following secretion from schistosome eggs it has to withstand the aggressive environment of the inflammatory egg granuloma characterized by low pH and high levels of oxidative and other denaturing agents.

The structural and mutational analysis of the complex of IPSEΔNLS and IgE revealed a positively charged interaction interface, which includes the IPSE/α-1-unique loop. Computational modeling suggested that the negatively charged region on the Cε2 domain of IgE provides a suitable binding interface for the positively charged IPSE/α-1 (Fig. 11, *A* and *B*).

We therefore developed a model for the mechanism of basophil activation by IPSE/α-1 (Fig. 11*C*), based on the computational modeling and our experimental data, specifically the inability of IPSE/α-1 to cross-link IgE, a putative 1:1 stoichiometry of IPSE/α-1·IgE complexes, and the binding of IPSE/α-1 to both Fab and Fc fragments: The first monomer of the IPSE/α-1 dimer binds the negatively charged surface of the Cε2 domain. Binding of the first IPSE/α-1 monomer to the IgE-Fc fragment facilitates binding of the second IPSE/α-1 monomer to the IgE-Fab. We hypothesize that this might trigger a conformational change of the IgE molecule in close proximity to the Fcε receptor and subsequent receptor activation without cross-linking. To validate this mechanism, further experiments are warranted to locate the binding site(s) of dimeric IPSE/α-1 on the IgE molecule by co-crystallization.

For the past 30 years, the receptor-mediated response to foreign antigens has been known to involve the high affinity IgE receptor on mast cells and basophils. Cross-linking of two or more IgE·receptor complexes on the cell surface stimulates degranulation of the cells and the release of mediators and cytokines (41). In contrast, IPSE/α-1 appeared to activate basophils by a novel mechanism that involves binding but not cross-linking receptor-bound IgE leaving the IgE "monomeric." A role of IgE monomers in initiating cell signaling upon binding to FcεRI

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

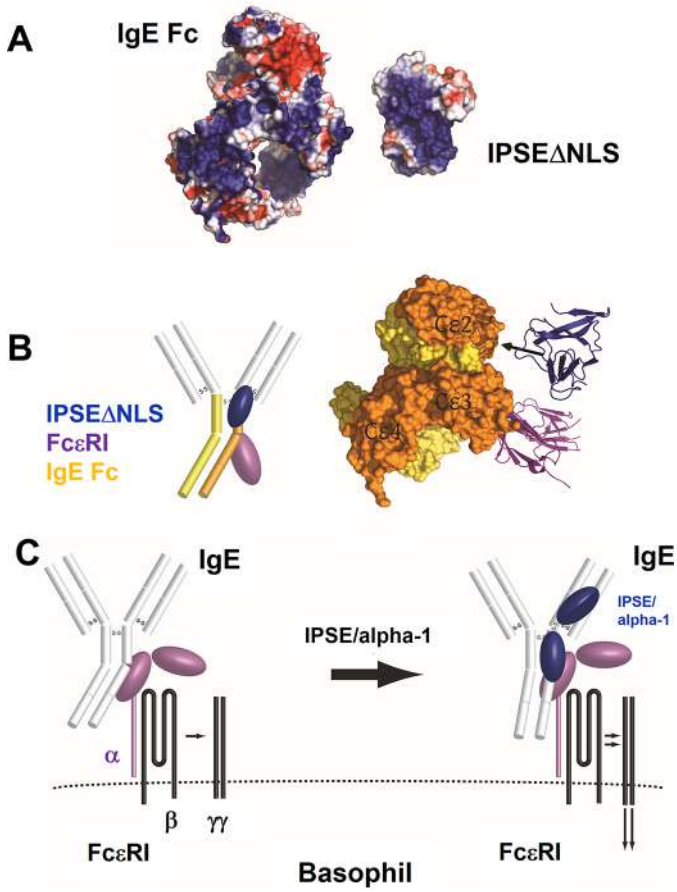


FIGURE 11. **Model of the interaction of IPSE/α-1 with IgE.** A putative novel mechanism of basophil activation is shown. *A,* surface representation of IgE Fc (PDB accession code 2Y7Q, *left*) and IPSEΔNLS (*right*) colored according to electrostatic surface potential at ±2 *kB T/e* for positive (*blue*) or negative (*red*) charge potential using the program APBS (48). *B,* proposed interaction between IPSEΔNLS and IgE (represented by the crystal structure of IgE Fc/FcεRI, PDB accession code 2WQR). The positively charged binding interface of IPSEΔNLS mapped by NMR and mutational analysis presumably interacts with a negatively charged surface on the Cε2 domain of IgE, while the FcεRI binds to the neighboring Cε3 domain of the IgE molecule. *C,* schematic model of IgE receptor activation on the surface of basophils. IPSE/α-1 may induce a structural rearrangement in the IgE molecule. This triggers the activation of the receptor signaling cascade.

has been controversially discussed (42, 43). High concentrations of monomeric IgE were reported to partially activate the FcεRI, thereby triggering cellular events such as degranulation and cytokine production (42). Moreover, different monoclonal IgEs seem to differ in their ability to activate basophils, but the structural differences that account for the functional divergence are not known (43). In contrast, Schweitzer-Stenner and Pecht consider nonphysiological experimental conditions to cause the effects of monomeric IgE (44). Recently, however, the cross-linking model for cell activation has been challenged also for the T and B cell receptors (45, 46). It was shown that inactive T and B cell receptors are present in an oligomeric form, whereas binding of their respective ligands leads to dissociation into monomeric receptors and subsequent clustering of the monomeric receptors (dissociation model) (45, 47). Further studies are necessary to determine the mechanism of action of IPSE/α-1. Because the solution methods used in this study do not necessarily reproduce the cell-surface environment in which IPSE/α-1 acts, further studies should reveal the structural rearrangement of FcεRI-bound IgE on the cell surface upon binding of IPSE/α-1.

In conclusion, the parasite-derived protein IPSE/α-1 is the first member of the βγ-crystallin superfamily discovered to have an immunomodulatory function. Additionally, this is the first report on the interaction of a crystallin fold with IgE. In contrast to other known IgE-dependent basophil activators, IPSE/α-1 does not cross-link IgE; instead it appears to activate basophils via a novel mechanism. Our results regarding the structure and the molecular function of the basophil IL-4-inducing factor IPSE/α-1 help to better understand the mechanism by which parasites modulate the host's immune system and may provide a starting point for pharmacological intervention in the future.

Author Contributions—N. H. M., H. H., M. S., and G. S. designed and coordinated this study and wrote the paper. N. H. M., K. T., T. M., and M. S, designed, performed, and analyzed NMR experiments, H. M. and J. M.-D. solved the crystal structure, and Si. B. performed and analyzed the experiments shown in Figs. 4, *A* and *B*, and 10, *A, B,* and *D*. D. B. provided technical assistance and performed experiments shown in Fig. 8. A. M. performed experiments shown in Fig. 5 and G. S. the experiments shown in Figs. 4*C* and 10*C*. S. B. and A. F. designed and performed the peptide scan (Fig. 7*D*). T. K., T. W., and G. S. designed, performed, and analyzed the SPR assays (Fig. 6). All authors reviewed the results and approved the final version of the manuscript.

Acknowledgments—We are grateful to Bernd Simon (EMBL Heidelberg, Germany), Gülden Demiraslan and Ulrike Schütz (Munich, Germany), Sven Mueller-Loennies and Niels Röckendorf (RCB, Borstel, Germany), and Prof. Brian J. Sutton and Prof. James M. McDonnell (King's College, London, UK) for support and discussions. We thank the Bavarian NMR Centre for NMR time, and the ESRF for access to measurements at beamline ID29. Editorial support was provided by David Young, Young Medical Communications and Consulting Ltd.

References

1. Schramm, G., and Haas, H. (2010) Th2 immune response against *Schistosoma mansoni* infection. *Microbes Infect.* **12,** 881–888
2. Seder, R. A., Paul, W. E., Davis, M. M., and Fazekas de St Groth, B. (1992) The presence of interleukin 4 during *in vitro* priming determines the lymphokine-producing potential of $CD4^+$ T cells from T cell receptor transgenic mice. *J. Exp. Med.* **176,** 1091–1098
3. Brunet, L. R., Finkelman, F. D., Cheever, A. W., Kopf, M. A., and Pearce, E. J. (1997) IL-4 protects against TNF-α 32 -mediated cachexia and death during acute schistosomiasis. *J. Immunol.* **159,** 777–785
4. Herbert, D. R., Orekov, T., Perkins, C., Rothenberg, M. E., and Finkelman, F. D. (2008) IL-4Rα expression by bone marrow-derived cells is necessary and sufficient for host protection against acute schistosomiasis. *J. Immunol.* **180,** 4948–4955
5. Sullivan, B. M., Liang, H. E., Bando, J. K., Wu, D., Cheng, L. E., McKerrow, J. K., Allen, C. D., and Locksley, R. M. (2011) Genetic analysis of basophil function *in vivo*. *Nat. Immunol.* **12,** 527–535
6. van Panhuys, N., Prout, M., Forbes, E., Min, B., Paul, W. E., and Le Gros, G. (2011) Basophils are the major producers of IL-4 during primary helminth infection. *J. Immunol.* **186,** 2719–2728
7. Schramm, G., Falcone, F. H., Gronow, A., Haisch, K., Mamat, U., Doenhoff, M. J., Oliveira, G., Galle, J., Dahinden, C. A., and Haas, H. (2003) Molecular characterization of an interleukin-4-inducing factor from

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

Schistosoma mansoni eggs. *J. Biol. Chem.* **278,** 18384–18392
8. Schramm, G., Mohrs, K., Wodrich, M., Doenhoff, M. J., Pearce, E. J., Haas, H., and Mohrs, M. (2007) Cutting edge: IPSE/α-1, a glycoprotein from Schistosoma mansoni eggs, induces IgE-dependent, antigen-independent IL-4 production by murine basophils *in vivo*. *J. Immunol.* **178,** 6023–6027
9. Schramm, G., Gronow, A., Knobloch, J., Wippersteg, V., Grevelding, C. G., Galle, J., Fuller, H., Stanley, R. G., Chiodini, P. L., Haas, H., and Doenhoff, M. J. (2006) IPSE/α-1: a major immunogenic component secreted from *Schistosoma mansoni* eggs. *Mol. Biochem. Parasitol.* **147,** 9–19
10. Dunne, D. W., Jones, F. M., and Doenhoff, M. J. (1991) The purification, characterization, serological activity and hepatotoxic properties of two cationic glycoproteins (α1 and ω1) from *Schistosoma mansoni* eggs. *Parasitology* **103,** 225–236
11. Young, N. D., Jex, A. R., Li, B., Liu, S., Yang, L., Xiong, Z., Li, Y., Cantacessi, C., Hall, R. S., Xu, X., Chen, F., Wu, X., Zerlotini, A., Oliveira, G., Hofmann, A., *et al.* (2012) Whole-genome sequence of *Schistosoma haematobium*. *Nat. Genet.* **44,** 221–225
12. Falcone, F. H., Dahinden, C. A., Gibbs, B. F., Noll, T., Amon, U., Hebestreit, H., Abrahamsen, O., Klaucke, J., Schlaak, M., and Haas, H. (1996) Human basophils release interleukin-4 after stimulation with *Schistosoma mansoni* egg antigen. *Eur. J. Immunol.* **26,** 1147–1155
13. Watanabe, T., Inoue, T., Ochi, H., Terashima, M., Asano, Y., and Nakatani, T. (1999) Lipid A directly inhibits IL-4 production by murine Th2 cells but does not inhibit IFN-γ production by Th1 cells. *Eur. J. Immunol.* **29,** 413–418
14. Patella, V., Florio, G., Petraroli, A., and Marone, G. (2000) HIV-1 gp120 induces IL-4 and IL-13 release from human FcεRI$^+$ cells through interaction with the VH3 region of IgE. *J. Immunol.* **164,** 589–595
15. Kaur, I., Schramm, G., Everts, B., Scholzen, T., Kindle, K. B., Beetz, C., Montiel-Duarte, C., Blindow, S., Jones, A. T., Haas, H., Stolnik, S., Heery, D. M., and Falcone, F. H. (2011) Interleukin-4-inducing principle from *Schistosoma mansoni* eggs contains a functional C-terminal nuclear localization signal necessary for nuclear translocation in mammalian cells but not for its uptake. *Infect. Immun.* **79,** 1779–1788
16. Wuhrer, M., Balog, C. I., Catalina, M. I., Jones, F. M., Schramm, G., Haas, H., Doenhoff, M. J., Dunne, D. W., Deelder, A. M., and Hokke, C. H. (2006) IPSE/α-1, a major secretory glycoprotein antigen from schistosome eggs, expresses the Lewis X motif on core-difucosylated *N*-glycans. *FEBS J.* **273,** 2276–2292
17. Mayerhofer, H., Schramm, G., Hatzopoulos, G. N., Mueller-Dieckmann, C., Haas, H., and Mueller-Dieckmann, J. (2009) Cloning, expression, purification, crystallization and preliminary x-ray crystallographic analysis of interleukin-4-inducing principle from *Schistosoma mansoni* eggs (IPSE/α-1). *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **65,** 594–596
18. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR.* **6,** 277–293
19. Sattler, M., Schleucher, J., and Griesinger, C. (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nuclear Magn. Reson. Spectrosc.* **34,** 93–158
20. Meyer, N. H., Schramm, G., and Sattler, M. (2011) (1)H, (13)C and (15)N chemical shift assignments of IPSEΔNLS. *Biomol. NMR Assign.* **5,** 225–227
21. Farrow, N. A., Muhandiram, R., Singer, A. U., Pascal, S. M., Kay, C. M., Gish, G., Shoelson, S. E., Pawson, T., Forman-Kay, J. D., and Kay, L. E. (1994) Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by 15N NMR relaxation. *Biochemistry* **33,** 5984–6003
22. Güntert, P. (2009) Automated structure determination from NMR spectra. *Eur. Biophys. J.* **38,** 129–143
23. Shen, Y., Delaglio, F., Cornilescu, G., and Bax, A. (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR.* **44,** 213–223
24. Linge, J. P., Williams, M. A., Spronk, C. A., Bonvin, A. M., and Nilges, M. (2003) Refinement of protein structures in explicit solvent. *Proteins* **50,** 496–506
25. Nederveen, A. J., Doreleijers, J. F., Vranken, W., Miller, Z., Spronk, C. A., Nabuurs, S. B., Güntert, P., Livny, M., Markley, J. L., Nilges, M., Ulrich, E. L., Kaptein, R., and Bonvin, A. M. (2005) RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. *Proteins* **59,** 662–672
26. Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR.* **8,** 477–486
27. Hooft, R. W., Vriend, G., Sander, C., and Abola, E. E. (1996) Errors in protein structures. *Nature.* **381,** 272
28. Riek, R., Wider, G., Pervushin, K., and Wüthrich, K. (1999) Polarization transfer by cross-correlated relaxation in solution NMR with very large molecules. *Proc. Natl. Acad. Sci. U.S.A.* **96,** 4918–4923
29. Otwinowski, Z., and Minor, W. (1997) Processing of x-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326
30. McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. (2007) Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674
31. Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F., and Vagin, A. A. (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **67,** 355–367
32. Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010) Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 486–501
33. Chen, V. B., Arendall, W. B., 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 12–21
34. Comeau, S. R., Gatchell, D. W., Vajda, S., and Camacho, C. J. (2004) ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Res.* **32,** W96–W99
35. Röckendorf, N., Bade, S., Hirst, T. R., Gorris, H. H., and Frey, A. (2007) Synthesis of a fluorescent ganglioside GM1 derivative and screening of a synthetic peptide library for GM1 binding sequence motifs. *Bioconjug. Chem.* **18,** 573–578
36. Haisch, K., Gibbs, B. F., Körber, H., Ernst, M., Grage-Griebenow, E., Schlaak, M., and Haas, H. (1999) Purification of morphologically and functionally intact human basophils to near homogeneity. *J. Immunol. Methods* **226,** 129–137
37. Mathieson, W., and Wilson, R. A. (2010) A comparative proteomic study of the undeveloped and developed Schistosoma mansoni egg and its contents: The miracidium, hatch fluid and secretions. *Int. J. Parasitol.* **40,** 617–628
38. Jaenicke, R., and Slingsby, C. (2001) Lens crystallins and their microbial homologs: structure, stability, and function. *Crit. Rev. Biochem. Mol. Biol.* **36,** 435–499
39. Hemmingsen, J. M., Gernert, K. M., Richardson, J. S., and Richardson, D. C. (1994) The tyrosine corner: a feature of most greek key β-barrel proteins. *Protein Sci.* **3,** 1927–1937
40. Smith, P., Fallon, R. E., Mangan, N. E., Walsh, C. M., Saraiva, M., Sayers, J. R., McKenzie, A. N., Alcami, A., and Fallon, P. G. (2005) *Schistosoma mansoni* secretes a chemokine binding protein with antiinflammatory activity. *J. Exp. Med.* **202,** 1319–1325
41. Holowka, D., Sil, D., Torigoe, C., and Baird, B. (2007) Insights into immunoglobulin E receptor signaling from structurally defined ligands. *Immunol. Rev.* **217,** 269–279
42. Kitaura, J., Xiao, W., Maeda-Yamamoto, M., Kawakami, Y., Lowell, C. A., and Kawakami, T. (2004) Early divergence of Fcε receptor I signals for receptor up-regulation and internalization from degranulation, cytokine production, and survival. *J. Immunol.* **173,** 4317–4323
43. Kitaura, J., Song, J., Tsai, M., Asai, K., Maeda-Yamamoto, M., Mocsai, A., Kawakami, Y., Liu, F. T., Lowell, C. A., Barisas, B. G., Galli, S. J., and Kawakami, T. (2003) Evidence that IgE molecules mediate a spectrum of

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

effects on mast cell survival and activation via aggregation of the FcεRI. *Proc. Natl. Acad. Sci. U.S.A.* **100,** 12911–12916
44. Schweitzer-Stenner, R., and Pecht, I. (2005) Death of a dogma or enforcing the artificial: monomeric IgE binding may initiate mast cell response by inducing its receptor aggregation. *J. Immunol.* **174,** 4461–4464
45. Yang, J., and Reth, M. (2010) Oligomeric organization of the B-cell antigen receptor on resting cells. *Nature* **467,** 465–469
46. Lillemeier, B. F., Mörtelmaier, M. A., Forstner, M. B., Huppa, J. B., Groves, J. T., and Davis, M. M. (2010) TCR and Lat are expressed on separate protein islands on T cell membranes and concatenate during activation. *Nat. Immunol.* **11,** 90–96
47. Yang, J., and Reth, M. (2010) The dissociation activation model of B cell antigen receptor triggering. *FEBS Lett.* **584,** 4872–4877
48. Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **98,** 10037–10041

Downloaded from http://www.jbc.org/ at MASARYKOVA UNIVERZITA on August 11, 2017

J Biomol NMR (2014) 59:135–145
DOI 10.1007/s10858-014-9832-4

ARTICLE

Robust and highly accurate automatic NOESY assignment and structure determination with Rosetta

Zaiyong Zhang · Justin Porter · Konstantinos Tripsianes · Oliver F. Lange

Received: 15 February 2014 / Accepted: 19 April 2014 / Published online: 21 May 2014
© Springer Science+Business Media Dordrecht 2014

Abstract We have developed a novel and robust approach for automatic and unsupervised simultaneous nuclear Overhauser effect (NOE) assignment and structure determination within the CS-Rosetta framework. Starting from unassigned peak lists and chemical shift assignments, autoNOE-Rosetta determines NOE cross-peak assignments and generates structural models. The approach tolerates incomplete and raw NOE peak lists as well as incomplete or partially incorrect chemical shift assignments, and its performance has been tested on 50 protein targets ranging from 50 to 200 residues in size. We find a significantly improved performance compared to established programs, particularly for larger proteins and for NOE data obtained on perdeuterated protein samples. X-ray crystallographic structures allowed comparison of Rosetta and conventional, PDB-deposited, NMR models in 20 of 50 test cases. The unsupervised autoNOE-Rosetta models were often of significantly higher accuracy than the corresponding expert-supervised NMR models deposited in the PDB. We also tested the method with unrefined peak lists and found that performance was nearly as good as for refined peak lists. Finally, demonstrating our method's remarkable robustness against problematic input data, we provided correct models for an incorrect PDB-deposited NMR solution structure.

Keywords Nuclear magnetic resonance · Automatic data analysis · Structure determination

Significance statement

Along with X-ray crystallography, NMR structure determination is the only available high-resolution method for protein structure determination. NMR spectroscopy is conducted in aqueous solution and thus be the only route towards high-resolution 3D structures for proteins that cannot be crystallized. However, the analysis of NMR data is very time-consuming and can generally only be conducted by highly trained NMR experts, which require between weeks and months to obtain accurate and precise structures. Here, we provide a novel method to automate the analysis process and show that accurate structures can be obtained. Remarkably, the automatically generated structures are, in a majority of the cases, more accurate than the structures laboriously generated by NMR experts. Our method promises to significantly increase the attractiveness and viability of NMR structure determination.

Introduction

Structure determination by nuclear magnetic resonance (NMR) spectroscopy is largely driven by distance

Electronic supplementary material The online version of this article (doi:10.1007/s10858-014-9832-4) contains supplementary material, which is available to authorized users.

Z. Zhang · J. Porter · O. F. Lange (✉)
Department Chemie, Biomolecular NMR and Munich Center for Integrated Protein Science, Technische Universität München, Lichtenbergstrasse 4, 85747 Garching, Germany
e-mail: oliver.lange@tum.de

K. Tripsianes
Central European Institute of Technology (CEITEC), Masaryk University, Kamenice 5, 62500 Brno, Czech Republic

O. F. Lange
Institute of Structural Biology, Helmholtz Zentrum München, Neuherberg, Germany

information gathered through nuclear Overhauser effect spectroscopy (NOESY). To use such data as distance restraints, the NOESY crosspeaks in multidimensional spectra have to be assigned to individual atoms of the biomolecular system. NOESY crosspeak assignment and structure generation steps are usually performed in an integrated, iterative manner. This maximizes the number of conformational restraints, while guaranteeing self-consistency amongst distance restraints (Wüthrich 1986).

Many of the repetitive tasks in NMR structure determination have been successfully automated (Guerry and Herrmann 2011; Moseley and Montelione 1999; Baran et al. 2004; Guntert 2009). Two such crucial tasks in the chain of the data analysis are the assignment of NOE crosspeaks and the determination of accurate structural models. Popular programs that perform these two tasks include ARIA (Linge et al. 2003a), CYANA (Guntert 2009; Guntert et al. 1997), AutoStructure (Huang et al. 2006) and UNIO (Serrano et al. 2012) and have recently been tested with good results in a blind-testing challenge (Rosato et al. 2012). However, a limitation of these programs is that they have to be able to generate a sufficiently accurate model from the initial set of assignments. This usually limits the methods to small proteins with high quality spectra, complete and accurate chemical shift assignments, and well-refined peak lists. When conditions are suboptimal, a calculation either does not converge, or worse, converges to a precise but inaccurate fold (Guerry and Herrmann 2011). Accordingly, these programs are not usually used unsupervised, and must instead be applied in combination with manual assignment and possibly peak list refinement by a skilled NMR expert. Indeed, in our own work on larger proteins, a few manual assignments were required to bootstrap the automated analysis with CYANA (Lange et al. 2012).

Here, we aim to develop an NOE assignment and structure determination algorithm that can—unsupervised—produce results that are both reliable and accurate. This algorithm should take chemical shift assignments and unassigned NOE peak lists as input and produce, without further user interaction, refined models of protein structures in atomic resolution.

To achieve this goal, we combine Rosetta structure prediction with automatic NOE assignment. It has been demonstrated that Rosetta, which searches for the lowest energy conformation of the polypeptide chain using physically realistic force fields, requires only very sparse NMR data to guide its search to accurate structures (Lange et al. 2012; Raman et al. 2010). The question we ask here is whether the very noisy automatically assigned NOE restraints might be able to provide sufficient guidance for Rosetta to yield accurate initial models. These models would then allow iterative refinement of NOE assignments until accurate high-quality structures and self-consistent assignments can be generated. Iteration of automatic NOE assignment with structural modeling is, however, also the basis of established algorithms. Thus, the crucial question to be explored in this study is not whether iteration between modeling and assignment is a successful strategy, but rather if a significant benefit is gained by using the improved, but computationally more demanding, ROSETTA structural modeling, and if we can solve the engineering challenge to render the ROSETTA structure calculation sufficiently robust against the very noisy automatically assigned NOE restraints of the initial assignment stage. In cases where established programs cannot find converged initial models, and thus fail, the new approach might converge and thus applicability is broadened to include more challenging cases. Additionally, the more accurate modeling provided by the ROSETTA energy function might render the method more robust against erroneous input data and yield more accurate final 3D models.

To couple NOE assignment with ROSETTA, we build on the previously developed iterative structural modeling algorithm, RASREC, and extend it to become an algorithm for automatic NOE assignment. This entails the implementation of a new ROSETTTA module for automatic NOE assignment as well as the development of a robust protocol to couple the iterative search for the near-native protein structures in RASREC with iterative NOE assignment. The assignment module employs among other techniques, network anchoring (Herrmann et al. 2002), ambiguous restraints (Nilges et al. 1997), covalent structure compliance (Herrmann et al. 2002; Huang et al. 2005), structure dependent and independent peak calibration, and restraint combination (Herrmann et al. 2002). In our final protocol, the calculation consists of multiple iterations of structural sampling guided by automatically assigned NOE restraints. In early iterations, cross-peak assignments compatible with preliminary models are reinforced, but incompatible assignments are not removed. In later iterations, incompatible cross-peak assignments are removed from the restraint list. Throughout the whole process, however, a pool of best fitting structures is maintained that is ranked by the initial NOE assignments. This is a major difference to existing programs and helps us to prevent convergence on inaccurate but self-consistent solutions. Implementation details of the new method will be described elsewhere (Lange 2013).

To investigate the performance of the new methodology, we carried out a benchmark on 50 NOE data sets obtained from 41 protein samples of 63–370 residues length. To test the impact of *difficult* inputs on the performance of auto-NOE-Rosetta, we have included unrefined and automatically picked peak lists, as well as sparse data sets obtained from perdeuterated ILV-methyl labelled protein samples. To avoid unwittingly cherry-picking targets that work especially well for our method, we chose three pre-existing

benchmark sets and used *all* monomeric proteins from each (Rosato et al. 2012; Lange et al. 2012; Mao et al. 2011).

In the following, we report on the results of the benchmark. First, we will contrast the performance of AutoNOE-Rosetta with CYANA. Subsequently, we compare the accuracy of the unsupervised method with the state of the art of expert guided NMR structure determination as reflected in PDB-deposited NMR models. This is followed by an analysis of structure validation metrics and NOE completeness scores. Finally, we stress test the method with non-ideal input data, such as raw or unrefined peak lists or incomplete and erroneous chemical shift assignments.

Results

Comparison with CYANA

We have defined a single set of parameters that is used to run all targets, including data preparation (e.g., automatic trimming of flexible tails), structure calculation and final model selection. Thus, results in similar quality as reported here should be achievable from application of the method to as yet unknown targets. We also provide a suite of scripts that allow the user to run the software in this unsupervised fashion. The entire benchmark set and the final models can be obtained from our website (www.csrosetta.org/benchmarks) and our results can be scrutinized by interested readers using our software and accompanying toolchain.

The benchmark comprises 50 NOE data sets derived from 40 different proteins ranging in size from 5.5 to 40 kDa. Input data are the sequence, chemical shift assignments and NOE peak lists ("Methods" section, SI Appendix, Table S1). In 20 cases, RDC data of the N–H bond vectors in one or more alignment medium was also included (SI Appendix, Table S1).

Multiple calculations are carried out with different weighting of the NOE data against the Rosetta Energy. One is selected from these based on a combination of final Rosetta Energy and the intrinsic precision of the resulting models ("Methods" section). Finally, to be accepted as a successful solution, the structures must fulfill two criteria: convergence and intrinsic NOE consistency (SI Appendix: Methods Section 2.2). AutoNOE-Rosetta was run successfully on 42 of 50 data sets, comprising 35 different proteins. Final models are shown in SI Appendix, Fig. S1 for all targets, and their accuracy is reported in SI Appendix, Table S2 as C_α-RMSD with respect to the reference structure. A number of targets have only been used after the autoNOE-algorithm was finalized, including all parameters, and the run selection protocol. These targets are DrR147D, MrR110B, OR8C, PfR13A, PsR293, SR384, SgR42, VpR247, and HmR11 and display similar performance as the other targets (SI Appendix, Table S2).

To provide a reference for the performance of autoNOE-Rosetta, we chose to run the popular program CYANA 3.0, which obtained the most accurate models in a recent community-wide blind structure determination challenge (CASD) (Rosato et al. 2012). In analogy to AutoNOE-Rosetta, we have defined an acceptance rule for CYANA. Based on suggestions of CYANA's creator, Peter Güntert, we use a combination of convergence and CYANA's target function (SI Appendix: Methods Section 2.1, Fig. S2).

CYANA was *successful* for 31 of 50 data sets according to its acceptance rule ("Methods" section). Thus, a significant improvement in both accuracy and radius of convergence for autoNOE-Rosetta is observed with respect to CYANA (Fig. 1). All structures that failed the automatic acceptance criteria in autoNOE-Rosetta also failed in CYANA, but eleven of the failing targets in CYANA were acceptable according to the criteria in autoNOE-Rosetta, and yielded accurate structures below 2.5 Å RMSD (Fig. 1). Furthermore, 10 of 17 inaccurate CYANA-structures (RMSDs $>$ 2.5 Å) were determined accurately by autoNOE-Rosetta (RMSDs $<$ 2.5 Å). Numerical values of the C_α-RMSD against the reference structures for CYANA and autoNOE-Rosetta can be found in (SI Appendix: Table S2).

Comparison with PDB-deposited NMR structures

The state-of-the-art in high-resolution NMR structure determination typically involves not just a single CYANA run, but performing several rounds of CYANA-based NOE assignment and refinement of the input peak lists (or even manual assignments, going through peak-by-peak), followed by simulated annealing in XPLOR or CNS (considered to have a better force field than CYANA), and finally a high-resolution refinement in explicit water (Linge et al. 2003b), where RDCs are used if present. To directly compare autoNOE-Rosetta to this more complex structure determination protocol, we included 20 protein targets in our benchmark for which both a conventionally determined solution NMR structure and an X-ray crystal structure are available. We further assume that the state-of-the-art in NMR structure calculation is well reflected in these 20 PDB-deposited NMR solution structures. Indeed, all these structures were deposited in the last decade, the program CNS is listed in all PDB headers (except 1xpv), and whenever the respective remark section is provided in the PDB header (12 of 20 cases), water refinement is mentioned explicitly.

In this study, we assume that the X-ray structure is an accurate representation of the dominant solution structure; accordingly, the RMSD of atomic coordinates between NMR and X-ray structure provides a measure for the accuracy of the NMR structure. This view is supported by the NMR data (SI Appendix: Table S3). Based on this criterion, autoNOE-Rosetta significantly outperforms conventional

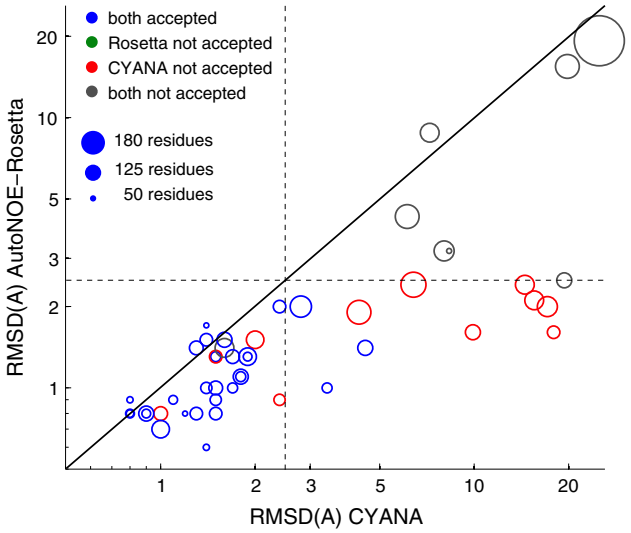


Fig. 1 Comparison of autoNOE-Rosetta with CYANA. Shown are the median C_α-RMSDs of final models with respect to their reference structure on logarithmic scale. The *diagonal line* indicates points of equal performance, points above the line correspond to targets for which CYANA yields lower RMSDs, and points below the line correspond to targets for which autoNOE-Rosetta yields lower RMSDs. The *dashed lines mark* 2.5 Å RMSD. The *size* of the proteins is proportional to the area of the symbol as indicated by the legend. The *color* indicates whether for CYANA, autoNOE-Rosetta or for both programs the final models are considered as success based on convergence and NOE consistency (SI Appendix, Method Section 2.0). RMSDs are capped at a maximum of 25 Å. Assignment statistics, convergence and accuracy of final models can be found in SI Appendix Table S7 and SI Appendix Table S8 for autoNOE-Rosetta and CYANA models, respectively. Comparing heavy-atom RMSDs instead of C_α-RMSDs yields a similar picture (SI Appendix, Fig. S6)

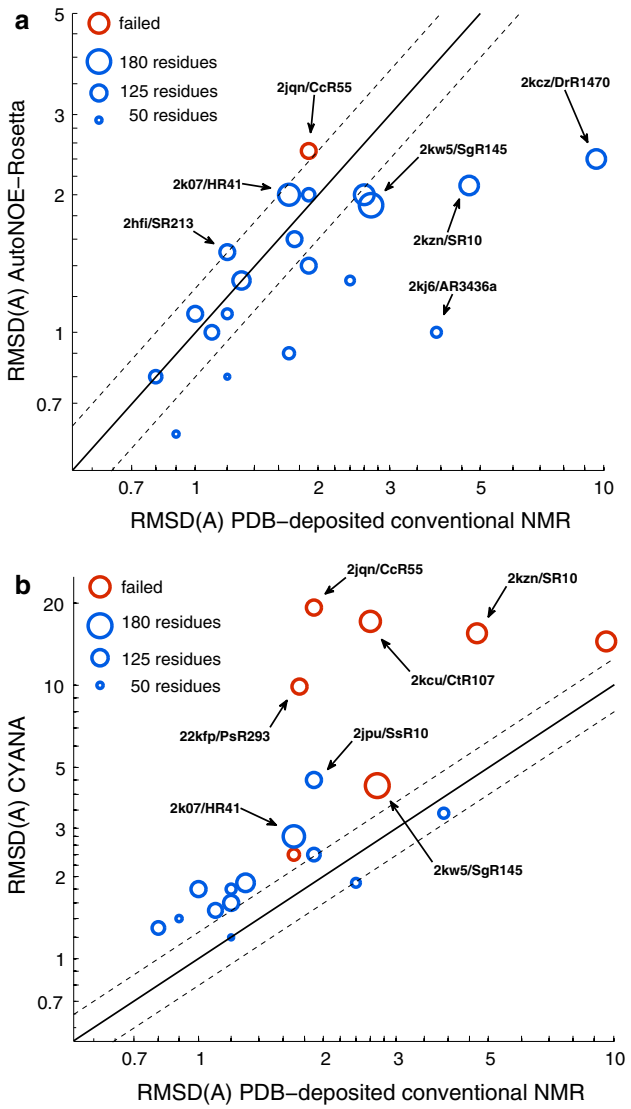


Fig. 2 Comparison of unsupervised automatic NOE models with expert-analyzed NMR solutions structures. The C_α-RMSDs of PDB deposited NMR models is plotted against final models obtained with **a** autoNOE-Rosetta and **b** CYANA. For AR3436a no X-ray structure is available as reference, but a new manually refined NMR solution structure, which supersedes 2kj6 (Fig. 4 and "Results" section). The *solid diagonal line* indicates points of equal performance, points above the line correspond to targets where PDB-deposited NMR structures have higher accuracy, and points below the line correspond to targets with higher accuracy of the autoNOE-Rosetta models. *Dashed lines mark* ±25 % accuracy. The *size* of the proteins is proportional to the area of the *symbol* as indicated by the legend. AutoNOE-Rosetta or CYANA runs that are not converged (<90 % of residues converged) are shown in *red*. Comparing heavy-atom RMSDs instead of C_α-RMSDs yields a similar picture (SI Appendix, Fig. S7)

supervised NMR structure determination (Fig. 2a and SI Appendix, Table S4). For 10 of 21 targets, accuracy is significantly improved, and only for 2 of 21 it is decreased (CcR55, partially converged; ER690 unconverged). Moreover, if we restrict the analysis to the 19 converged targets, accuracy never deteriorates more than 33 %, whereas it improves for 7 targets significantly beyond 33 %. This is in stark contrast to the performance of established automatic assignment programs. Only 3 of the smallest targets of the benchmark set (<80 residues) yield sufficiently accurate results in CYANA to compete with PDB deposited NMR structures. For the other 18 of 21 targets, the structures obtained unsupervised with CYANA are >25 % worse in accuracy than PDB-deposited NMR structures (Fig. 2b). Of these 18 with deteriorated accuracy, 13 yield a tight structural bundle and 10 are acceptable according to the success criteria introduced above (SI Appendix: Tables S2–S8).

Analysis of structural quality

In addition to a high accuracy, we generally would like to obtain 3D models of proteins with a high structural quality. This quality is generally assessed by structural validation packages through various metrics, such as packing quality, Ramachandran consistency, and Janin-plots. NMR solution structures based on NOE distance restraints are prone to show deficits (Doreleijers et al. 2011), whereas un-restrained CS-Rosetta models were previously reported to show high

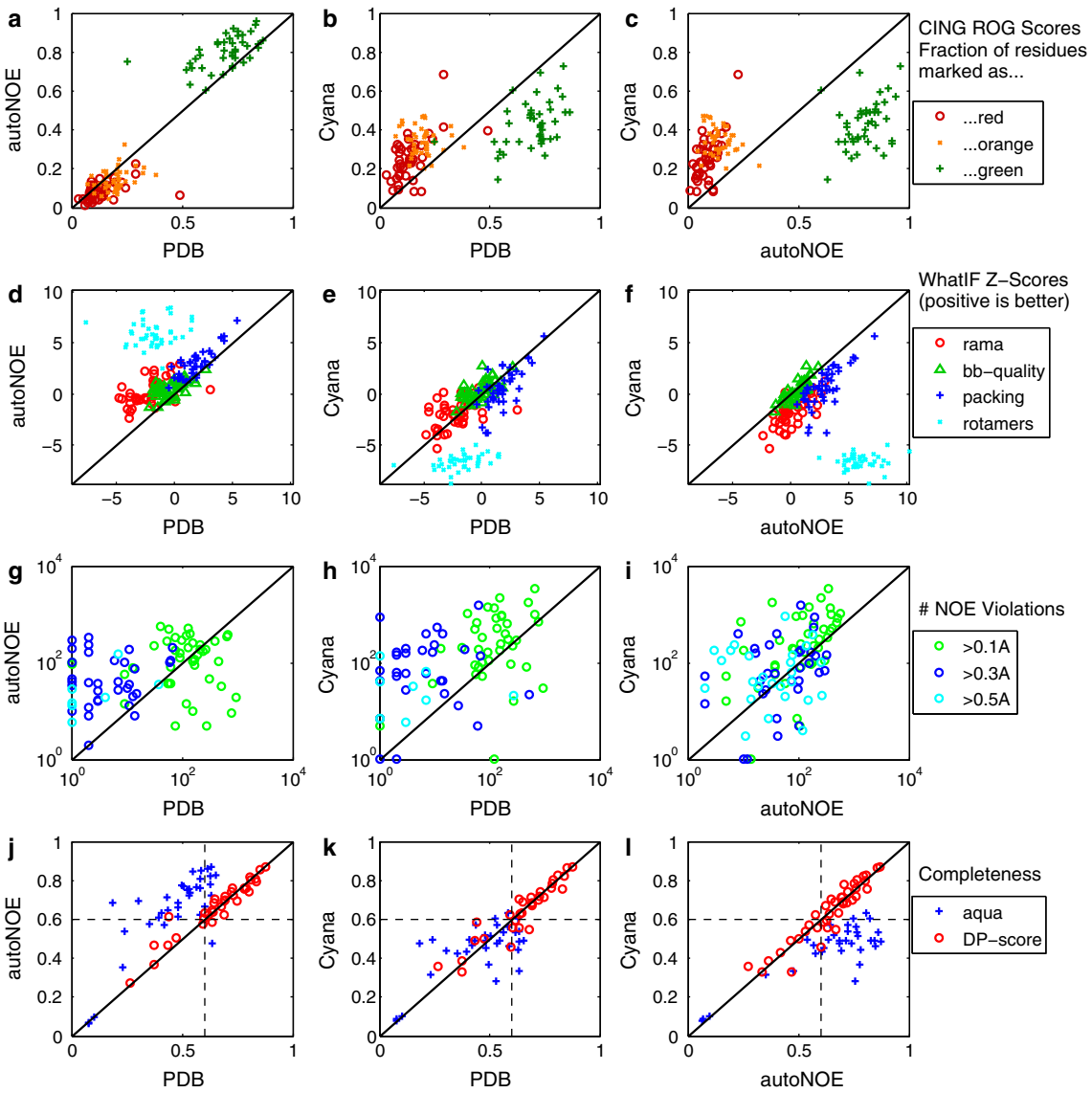


Fig. 3 Validation metrics for autoNOE-Rosetta, CYANA and PDB-deposited NMR models. Metrics computed for autoNOE-Rosetta and CYANA-models are compared to metrics computed on PDB-models, in panel-columns 1 and 2, respectively. Metrics between autoNOE and CYANA are conmpared directly in panel-column 3. **a–c** Fraction of residues annotated as *red, orange* and *green* by the iCING server's ROG score(legend). Less *red* and *orange* and more *green* residues is better. **d–f** WhatIF Z-scores for Ramachandran plot appearance, backbone-quality, packing and chi-1/chi-2 rotamer normality (legend). Higher Z-scores are better. **g–i** The number of NOE restraints violated by structural models. Structural models of CYANA and autoNOE are analyzed together with the restraints produced by the respective algorithms. PDB-deposited models are analyzed with respect to the NOE restraints uploaded with the structures. **j–l** Completeness scores computed with AQUA (Doreleijers et al. 1999) and AutoStruct-DP (Huang et al. 2005). Higher numbers are better

structural quality but significantly lower accuracy than NOE-driven structure calculations (Rosato et al. 2012). We were curious to see whether autoNOE-Rosetta both, preserves the high structural quality, despite being subjected to a large number of automatically assigned NOE restraints, and yields more accurate structures than CS-Rosetta. To assess the structural quality of AutoNOE-Rosetta models, we used the online validation server iCING (Doreleijers et al. 2012), which performs WhatIF (Vriend 1990), PROCHECK (Laskowski et al. 1996) and its own structural analysis.

The iCING-ROG score summarizes and integrates different validation measures into a single score and annotates individual residues as *green*, *orange* and *red* to convey an increasing level of alertness for unphysical local structure (Doreleijers et al. 2012). AutoNOE-Rosetta models produce generally less red and orange residues than PDB NMR-models or CYANA models (Fig. 3a–c). WhatIF compares local structure of the protein against common structural knowledge derived from high-resolution X-ray structures (Vriend 1990). Figure 3d–f shows the WhatIF

structure Z-scores on *Ramachandran plot appearance, backbone quality, 1st generation packing quality,* and *chi-1/chi-2 rotamer quality*. AutoNOE-Rosetta models generally are of higher quality than PDB NMR models or CYANA models.

Analysis of restraint violations

Another popular criterion for judging NMR structure quality is a low count of restraint violations by the final models. Figure 3g–i shows how often the final models violate the NOE-derived restraints by >0.1, >0.3 and >0.5 Å. Generally, NMR restraint-sets deposited with their corresponding PDB structures have less violations above >0.3 or >0.5 Å than those obtained with CYANA or autoNOE-Rosetta, but CYANA and AutoNOE-Rosetta yield similar results. We found that, for AutoNOE-Rosetta ensembles, many of the violations occurred at side-chains that adopted multiple conformations. In these cases, each conformation would actually be consistent with a subset of the violated NOE restraints involving this side-chain, and it would be plausible that dynamic averaging causes the assigned NOE cross-peaks. Since it is well possible that dynamic averaging might be the reason for some of the observed violations, as well as the fact that programs could trivially remove any violated restraint from the restraint-list, it is questionable whether the count of restraint violations is actually a valuable criterion for NMR structure validation. Indeed, we see no particular correlation between this measure and accuracy of the final models (C_α-RMSD) regardless whether they were downloaded from the PDB or generated with CYANA or autoNOE-Rosetta (SI Appendix: Fig. S3).

Analysis of NOE completeness

Since the autoNOE-Rosetta structures fit more accurately to X-ray structural models, a possible concern might be that Rosetta modeling is biased towards X-ray crystallographic artifacts rather than solution state structure. To verify that this is not the case, we show that autoNOE-Rosetta models yield a better or equivalent interpretation of the NMR data in comparison to conventional NMR solution structures, as quantified by the AQUA completeness (Doreleijers et al. 1999) and the AutoStruct DP score (Huang et al. 2005). AQUA reads the models and restraint list and checks how many of the proton–proton contacts in the model are actually observed as assigned NOEs. The more modern DP score uses chemical shift assignments and unassigned peak lists as input, and is thus independent of the specific restraint list. AQUA's completeness score is systematically better for AutoNOE-Rosetta than for PDB-NMR or CYANA models (Fig. 3j–l, black circles). For most targets the DP scores are comparable between the different methods (Fig. 3j–l, blue crosses). However, for some PDB NMR structures with low DP-scores (<0.6), autoNOE-Rosetta was able to yield significant improvements. Overall these quality measures show that autoNOE-Rosetta models yield an interpretation of the NMR data that is as good or better than that of the PDB-deposited NMR models.

Performance under non-ideal input

Next, we were interested how autoNOE-Rosetta behaves when provided with problematic data. We tested both automatic (raw) and refined peak lists for 8 targets from round II of the blind, community-wide NMR structure determination challenge (CASD) (Rosato et al. 2009). In addition to the 8 *raw* data sets, we use 7 *unrefined* data sets from previous work (Lange et al. 2012) and one from CASD round I. For these *unrefined* data sets, peaks have been picked manually and chemical shift assignments have been validated, but the peak lists and chemical shift assignments have not yet undergone iterative refinement using structural models. Of the *unrefined* data sets, 6 stem from ILV-methyl labeled perdeuterated protein samples. Restraints obtained from such ILV-samples are inherently sparse, rendering structure calculation more challenging due to a lower restraint density. Moreover, the sparser NOE networks render the automatic validation of NOE cross-peak assignments via network anchoring less effective.

The availability of 9 targets with both *raw/unrefined* and *refined* data allows us to investigate the robustness of autoNOE-Rosetta. AutoNOE-Rosetta turns out to be remarkably robust; for 7 of the 9 *raw/unrefined* peak lists differences in accuracy are insignificant (<0.3 Å). In only two cases, StT322 and HR5460, was the accuracy significantly decreased. The automatic acceptance criteria successfully identified both these raw data sets as having produced untrustworthy results. Interestingly, autoNOE-Rosetta tends to select a lower weight for the NOE-based pseudo-energy contribution for *raw* peak lists compared to *refined* peak lists (Table 1), which is consistent with the presumed lower quality of the data.

Another type of challenging input is given by the 8 *unrefined* ILV data sets. AutoNOE-Rosetta succeeded on four of these data sets and yielded a partially converged structure for another (HmR11). CYANA, however, did not succeed on any of these 8 data sets. Is the deciding factor, which makes these data sets so challenging, the sparseness of the ILV data, the quality of the data sets (*unrefined* vs. *refined*), or the increased molecular weight (ILV data sets have a molecular weight between 15 and 21 kDa)? One can mostly exclude the increased molecular weight, as the driving factor for these failures, since both autoNOE and CYANA were significantly more successful on the *refined*

Table 1 Impact of raw peak lists

Target	Reference	Size	Residue ranges			Weight ratio[a]	C_α-RMSD (Å) to reference structure			
			of NMR sample	Used in Rosetta	RMSD analysis		Raw peak list		Refined peak list	
							CYANA	autoNOE-Rosetta	CYANA	autoNOE-Rosetta
StT322	2loj	38	1–63	26–63	26–63	0.04	8.3	3.2	1.4	1.7
HR6470	2l9r	48	1–69	11–58	11–58	1.00	0.8	0.9	0.8	0.8
OR135	2ln3	69	1–79	5–73	5–73	0.50	0.9	0.8	1.1	0.9
AR3436a	tbd[c]	80	1–97	14–93	14–93	0.40	3.4	1.0	1.7	1.0
HR6430	2la6	89	1–99	11–99	11–99	1.00	1.4	1.0	1.5	0.9
HR2876	2ltm	95	1–107	13–107	13–107	0.04	Not converged	1.6	1.4	1.5
YR313	2ltl	102	1–119	18–119	18–40, 46–115	0.10	1.6	1.4	1.7	1.3
OR36	2lci	128	1–134	1–128	1–128	0.50	2.0	1.5	n/a[b]	1.2
HR5460	2lah	150	1–160	11–160	19–160	0.02	Not converged	3.2	n/a[b]	1.8

[a] Ratio of NOE-pseudo energy weights selected automatically by autoNOE-Rosetta for *raw* versus *refined* data sets

[b] A segmentation fault in CYANA 3.0 prohibited us from finishing the structure calculation

[c] The reference structure for AR3436a is a new NMR structure (PDB accession code: TBD) that results from a (manual) re-evaluation of the original NMR spectra

data sets in the same size range. Furthermore, we showed above that the influence of data quality (*raw* vs. *refined)* on autoNOE-Rosetta is low for small, double-labeled data sets. Thus, the lower success rate is likely a result of the sparseness of the ILV data.

Detecting problematic or unsuccessful calculations

To run autoNOE-Rosetta or CYANA unsupervised, it is important to have clear criteria to flag problematic runs. This filter mechanism has to catch most, if not all, problematic results. In other words, the filter should produce little or no false positives. Some false negatives, on the other hand, are not as worrisome, as human experts can inspect a few such calculations. Here, we have introduced clear definitions for such a filter rule based on convergence of structures and NOE self-consistency for CYANA and autoNOE-Rosetta (SI Appendix: Methods Section 2). Of the eight declined calculations performed with autoNOE-Rosetta, four failed both criteria, and four (two each) failed only one of the criteria. The data sets that only failed the consistency criterion, are YR313(raw) and StT322(raw). While YR313 yielded accurate structures (1.4 Å) in autoNOE-Rosetta, StT322 did not (C_α-RMSD 3.2 Å; SI Appendix Fig. S4). For CcR55, HR5460(raw), and HmR11(unrefined) only 88, 79 and 75 % of residues converged, respectively, failing the criterion of 90 % convergence by only a small margin. In these cases, the converged part of the structure is reasonably accurate (C_α-RMSDs of 1.3, 2.0, and 3.1 Å, respectively) and would provide an advanced starting point for further iterative and structure based refinement of the data set (SI Appendix, Fig. S4). Hence, as intended, the filter has been successful in producing no false positives and only very few false negatives.

Detailed discussion of target AR3436a

The data presented here shows that autoNOE-Rosetta yields accurate results even when the peak lists are not well refined. In the following we discuss a fortuitous discovery that demonstrates that autoNOE-Rosetta is not only robust against problematic peak lists, but also shows remarkable accuracy in the face of incomplete or erroneous side-chain chemical shift assignments. During our work on the here-presented benchmark we were initially puzzled by one outlier. For this outlier, AR3436a, autoNOE-Rosetta yielded structures that were 3.8 Å away from the PDB-deposited NMR solution structure (2kj6). The AR3436a data set stems from the CASD set, and was originally posed as a blind challenge to the community. The results of this competition seemed fairly standard except the CS-Rosetta models were identified as an outlier (Rosato et al. 2012): all NOE driven programs produced structures close to the PDB-deposited structure (1.4–2.2 Å) and with acceptable, albeit slightly borderline, validation scores. However, a closer inspection of the NMR solution models (2kj6) reveals that the main helix is at an angle causing the hydrophobic core of the protein to be exposed (Fig. 4a–c). In the autoNOE-Rosetta models, in contrast, the helix is well packed against the core (Fig. 4b–d), which is more consistent with our understanding of the physical chemistry of hydrophobic protein cores. Moreover, the CS-Rosetta based submissions to the blind structure determination challenge also packed the helix against the core (with

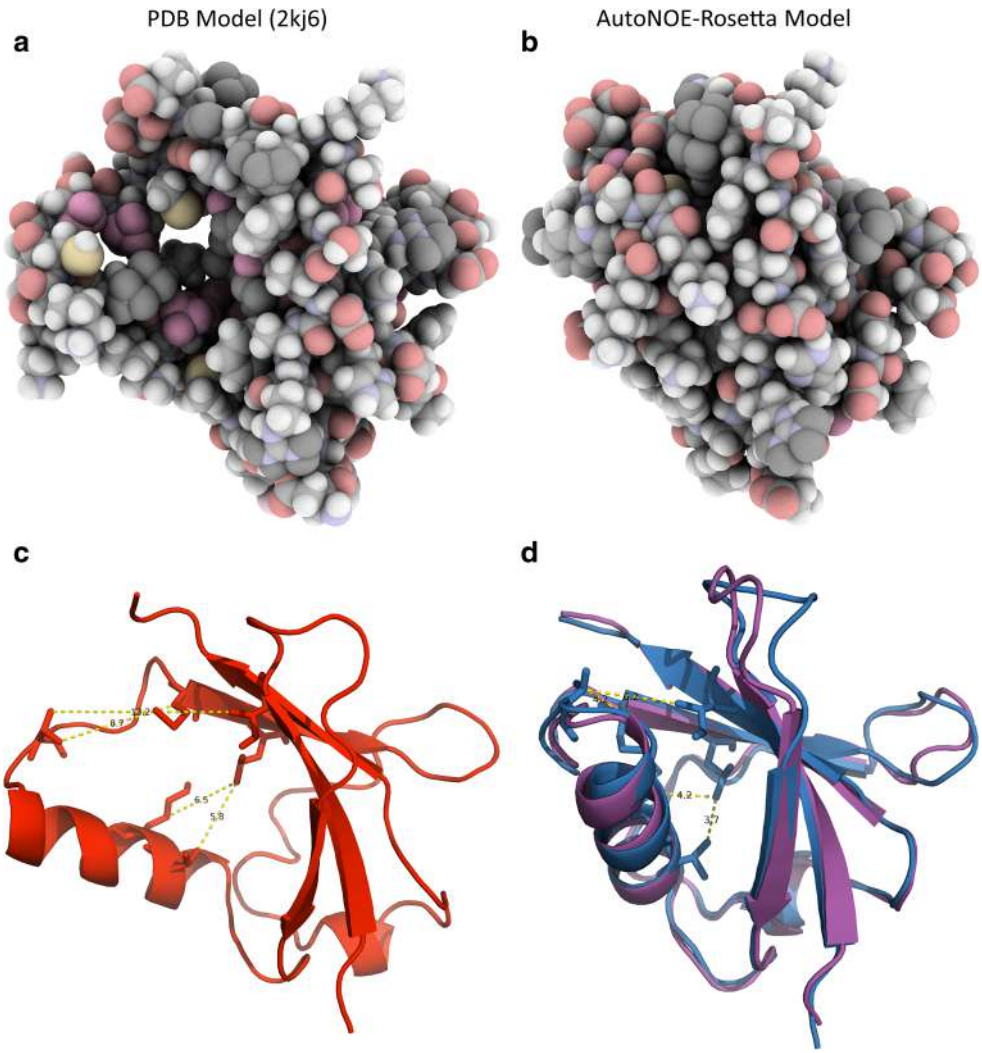


Fig. 4 Structure determination of AR3436A from incomplete and erroneous input data. Shown are two models of AR3436A in space-fill (**a**, **b**) and cartoon visualization (**c**, **d**) to highlight the differences in packing of the hydrophobic core between the PDB-deposited NMR solution structure (**a**, **c**) and the structure obtained with AutoNOE-Rosetta from the same input data (**b**, **d**). Due to the incomplete and erroneous chemical shift assignments AutoNOE-Rosetta can only assign a few NOE-crosspeaks (*yellow lines*) that support the packing of the helix, nethertheless, these are sufficient to yield well packed structures. **d** The PDB-deposited models violate these NOE crosspeaks, demonstrating that the respective assignments were discarded because they didn't fit initial models

RMSD > 4 Å to the reference NMR structure), but did not converge to a high-precision structural bundle.

These observations prompted us to investigate whether the better-packed structure obtained with AutoNOE-Rosetta might actually be better supported by the raw NMR data as well. Indeed, a careful analysis of the raw input data conducted together with members of the laboratory that authored the original data set revealed a number of problems[manuscript in preparation]. Although the backbone assignment was nearly complete and correct, the side-chain chemical shifts were incomplete and had miss-assignments. Additionally, the NOESY data were under-picked as indicated by the unbalanced Recall-Precision scores of the PSVS analysis, such that many potentially well resolved peaks were not contained in the original peak list. These issues hindered the structure calculations of NOE-driven programs, but had no influence on the CS-Rosetta calculations. After correcting these issues with the input data, the structures obtained with conventional methods matched with the autoNOE-Rosetta models obtained with either the original data (1.0 Å) or the new data (1.0 Å). This shows that autoNOE-Rosetta is not only reliable with unrefined (raw) peak lists but also with raw (i.e., incomplete and erroneous) sidechain chemical shift assignments. We are now in the process of systematically investigating the influence of such raw chemical shift assignments on automatic NOE assignment methods and our preliminary results support the anecdotal case reported here. The advantage of this robustness of AutoNOE-Rosetta for the full NMR pipeline is obvious. Assignment of side-chain chemical shifts is often a major bottleneck to progress in an NMR structure determination project. Automatic methods, such as FLYA (Schmidt and Guntert 2012), might take the burden of manual assignment, but cannot be relied on to always yield the highest quality of resonance assignments. However, paired with autoNOE-Rosetta, which is more fault-tolerant than other methods, an accurate structure might still be generated either as final result, or as a starting point for further refinement of the chemical shift assignments.

Discussion

We developed a new method for automatic NOE assignment and NMR structure determination, which we tested on a benchmark of 50 data sets including 20 for which X-ray crystallographic reference structures were available. A final convergence and NOE consistency filter accurately discriminates between successful and failed runs, and all 42 runs that pass this filter yield an accuracy better than 2.5 Å C_α-RMSD. Thus, we successfully combined the most important traits of CS-Rosetta with those of NOE-driven structure determination. The new algorithm is robust against missing or erroneous data as CS-Rosetta, but in the end exploits the full NOESY data to achieve the optimal precision and accuracy in final structures. In particular the lack of precision is problematic for CS-Rosetta, even if NOE-based filtering is applied (CS-DP-Rosetta (Raman et al. 2010)), as shown by the community wide assessment of structure determination (CASD) (Rosato et al. 2012).

The usefulness of an automatic NOESY assignment algorithms hinges on its ability to handle a wide variety of data. In fact, the quality of NOESY peak lists can vary dramatically as a function of the quality of the raw data, the method of picking peaks, and the level of peak list refinement. With 50 data sets from 41 different proteins, we are confident that our benchmark covers a realistic range of NMR data quality. To enhance the variety in the benchmark, we also included data sets at different stages of refinement (termed *raw*, *unrefined*, and *refined*). And in spite of this wide variety of input data quality autoNOE-Rosetta yields accurate results with striking consistency, which demonstrates a remarkable robustness of the method against challenging input data. Thus, autoNOE-Rosetta is a significant advance in *fully automatic* analysis of NMR data.

We were able to compare autoNOE-Rosetta ensembles with PDB-deposited NMR ensembles which reflect the state-of-the-art in NMR structure determination including final refinement in explicit water. Remarkably, the autoNOE-Rosetta results are either very close in accuracy (within 25 %) or significantly better (Fig. 2) than the PDB-deposited models. The most significant improvements were from 9.6 to 2.3 Å for the double-labelled sample, DrR1470, and from 4.7 to 2.1 Å for the triple-labelled, ILV-protonated sample, SR10 for which our calculations started from an *unrefined* data set.

AutoNOE-Rosetta ensembles' high accuracy—both relative and absolute—is especially remarkable considering that we are comparing an automated, unsupervised method with expert driven iterative and structure based refinement, as it is reflected in PDB deposited structures. For experts in NMR data analysis the method will provide better starting points for refining challenging data sets. For non-experts it will allow a safe and straightforward application of NMR structure determination to routine cases. We are confident that our method provides a significant progress towards unsupervised automatic NMR structure determination, which is likely to broaden the applicability of NMR for structure determination in academic and non-academic labs.

Methods

Benchmark

The 50 data sets comprising target sequence, assigned chemical shifts, and unassigned peak lists were obtained from three published sources (SI Appendix: Table S1): (1) all data sets available by December 2012 at the community wide assessment of NMR structure determination (CASD) (Rosato et al. 2012; Rosato et al. 2009) (currently hosted at http://www.wenmr.eu/wenmr/casd-nmr-data-sets), (2) all monomer data sets from a recent molecular replacement (MR) benchmark (Mao et al. 2011) (http://psvs-1_4-dev.nesg.org/MR/dataset.html) (SI Appendix: Fig. S5), (3) all targets from our previous work (Lange et al. 2012).

Peak lists from the first prediction period of CASD (CASDI) are refined. For targets from the second prediction period of CASD (CASDII), both, *refined* and *raw* (automatically picked) peak lists are available. For MR targets, the status of the peak lists is unknown but assumed *refined*, and for ILV-targets the peak lists and chemical shift files are *unrefined*, that is chemical shift assignments have been verified and peaks have been picked by a human expert (Lange et al. 2012), but the data sets have not undergone iterative refinement using structural models.

To analyze the accuracy of final structures, we computed the C_α-RMSD on all residues that are structured in the reference. Tails that were not well defined (flexible) in the reference structure are excluded from RMSD computation as specified in Table S1. For 11 reference structures, also internal loop-regions were not well defined and had to be excluded from RMSD calculations. Detailed justifications for these exclusions are given in Table S9. For a given method, autoNOE-Rosetta or CYANA, the ten final models are superimposed with the reference structures to compute C_α-RMSDs and heavy-atom RMSDs.

AutoNOE-Rosetta

AutoNOE-Rosetta structure calculations were run with parameters as detailed here (Lange 2013). Fragments were picked by the Rosetta3 fragment picker (Vernon et al. 2013) using the provided chemical shift data. Homologous proteins using an e-value cutoff of 0.05 (sequence identity

>20 %) were excluded from fragment picking. Tolerances for NOESY cross-peak assignment were set for all targets to 0.3, 0.3, 0.03 and 0.04 for ^{13}C, ^{15}N, direct ^{1}H, and indirect ^{1}H dimension, respectively. Residual Dipolar Coupling data were used where available (SI Appendix, Table S1).

For data sets with *unrefined* or *refined* peak lists, NOE-restraint strengths of 5, 10, 25 and 50, respectively are chosen, and for targets with *raw* peak lists restraint strengths of 1, 2, 5, 10, 25, and 50. For each restraint weight 3 independent runs were carried out with different random seeds. The 10 lowest energy structures yield the final ensemble of a given run.

To identify the optimal run the resulting ensembles were ranked as follows: The converged residues are identified as those with a C_α-RMS fluctuation of less than 2 Å, as reported previously (Lange et al. 2012). The average pairwise RMSD is computed on converged regions (SI Appendix, Methods), and an effective precision (EP) is computed from pairwise RMSD and fraction of converged residues. For each run with constraint weight w_{cst} a cumulative score $S = E - 12 \log w_{\mathrm{cst}} + 5\mathrm{EP}$ is computed, where E denotes the median Rosetta all-atom energy of the ensemble. If in any of the runs more than 2,000 peaks with initial assignments are removed, because final models violate them (column 'violated' in Table S11), only E is considered for selection of runs, otherwise the final run is selected using S.

The models of the top-ranking ensemble are further relaxed against the automatically assigned NOEs including intra-residue and sequential NOEs using a tenfold increased NOE-restraint weight. If this procedure reduces the number of NOE violations to less than 40 % of the violations counted in the ensemble of un-relaxed models, the relaxed models are accepted as final models, otherwise the un-relaxed models are kept as final models. This was the case for data sets HR2876(raw), YR313(raw), and CtR107. For all other data sets the relaxed models are kept as final models. This refinement step generally reduces NOE-violations without significantly affecting backbone RMSD to the reference structure.

We established two criteria for *successful* calculations: (1) reasonable NOE consistency (target-function <500) and (2) convergence (SI Appendix: Methods 2.2). For the convergence criterion the number of well defined residues has to reach 90 % or more of the total number of residues with random coil index (RCI) derived S (Guerry and Herrmann 2011) order parameter (Berjanskii and Wishart 2005) larger than 0.7 ("Methods" section and SI Appendix, Table S2d).

A few NOE data sets were recorded with reduced sweep width leading to *peak folding*. AutoNOE-Rosetta unfolds such frequencies on the fly, if the sweep-window is noted in the header of the respective peak list. For CYANA calculations we manually *unfolded* by replicating peaks with integer multiples of the sweep width subtracted or added to the respective frequencies. This applies to four peak lists of two proteins of our benchmark and the corresponding sweep-width parameters are given in (SI Appendix, Table S6).

AutoNOE-Rosetta is parallelized for the MPI framework and runs were either carried out on our in-house cluster or on JUROPA at the Juelich Supercomputer center using 184 or 192 parallel processes, respectively.

Instructions to run autoNOE-Rosetta including command-lines can be found in the Manual or Tutorial sections of our website (www.csrosetta.org) and in SI Appendix: Methods.

Cyana structure calculations

Cyana 3.0 calculations were carried out to provide readers with a familiar reference for each target. TALOS+ restraints were generated from the chemical shift data, and 100 initial, and 20 final models were generated using 20,000 steps of torsion angle dynamics. RMSDs were computed from the 20 final models using the same residues and reference structure as for autoNOE-Rosetta models (SI Appendix: Methods for example script). All TALOS+ predicted phi and psi angels with prediction class 'Good' are used. Two schemes to derive torsion restraints from TALOS+ predictions were tested. ACO_TIGHT restraints were generated by computing the lower- and upper bound as $\phi \pm \Delta\phi$, where ϕ denotes the TALOS+ predicted torsion angle in degree, and $\Delta\phi$ the TALOS+ estimated standard deviation. For ACO_LOOSE, we obtained bounds as $\phi \pm 2\max(\min(\Delta\phi, 35), 10)$. ACO_TIGHT is the recommended protocol at the NMR facility of the Center for Advanced Biotechnology and Medicine (CABM) as described here (http://www.nmr2.buffalo.edu/enter/NMRWiki/images/2/2e/Talos2dyana_taloserrors.txt). ACO_LOOSE is the protocol that derives from applying the talos2dyana.com executable packaged with the TALOS+ software. A comparison of both protocols shows that ACO_TIGHT yields better accuracy over all targets (SI Appendix, Fig. S2a). Thus, ACO_TIGHT is used in all further CYANA calculations.

Where RDC data was available, CYANA runs were carried out both, with and without RDC data. A weight of 0.02 was used for the RDC restraint, and 0.2 as cutoff for RDC violation output. For each alignment medium 5 additional pseudo-residues of type LL5 and 1 of type ORI are attached at the end of the protein sequence. Alignment tensor parameters, Dzz and R, are estimated using the macro FindTensor.cya which employs the histogram method (Clore et al. 1998). This protocol was obtained from http://

www.nmr2.buffalo.edu/nesg.wiki/CYANA. RMSDs of CYANA calculations with RDCs were generally higher than CYANA calculations without RDCs (SI Appendix, Fig. S2b), whereas RDC data leads to improved results for autoNOE-Rosetta (SI Appendix, Fig. S2c). Thus, CYANA calculations *without* RDCs are compared to autoNOE-Rosetta *with* RDCs throughout the study.

Acknowledgments We are indebted to the work of the Northeast Structural Genomics Consortium, which has made available NMR data sets for such a high number of proteins including those with known Xray structure. Only the extensive high-quality work of the NESG and its open attitude towards sharing data enabled this study in the breadth and thoroughness now presented. Furthermore, we have to thank Peter Güntert and Paolo Rossi for many helpful discussions, as well as Michael Sattler for carefully reading the manuscript. This work was supported by DFG grant LA 1817/3-1.

References

Baran MC, Huang YJ, Moseley HNB, Montelione GT (2004) Automated analysis of protein NMR assignments and structures. Chem Rev 104:3541–3556

Berjanskii M, Wishart D (2005) A simple method to predict protein flexibility using secondary chemical shifts. J Am Chem Soc 127:14970–14971

Clore GM, Gronenborn AM, Bax A (1998) A robust method for determining the magnitude of the fully asymmetric alignment tensor of oriented macromolecules in the absence of structural information. J Magn Reson 133:216–221

Doreleijers JF, Raves ML, Rullmann T, Kaptein R (1999) Completeness of NOEs in protein structures: a statistical analysis of NMR data. J Biomol NMR 14:123–132

Doreleijers JF et al (2011) NRG-CING: integrated validation reports of remediated experimental biomolecular NMR data and coordinates in wwPDB. Nucleic Acids Res 40:D519–D524

Doreleijers JF et al (2012) CING: an integrated residue-based structure validation program suite. J Biomol NMR 54:267–283

Guerry P, Herrmann T (2011) Advances in automated NMR protein structure determination. Quart Rev Biophys 44:1–53

Guntert P (2009) Automated structure determination from NMR spectra. Eur Biophys J 38:129–143

Guntert P, Mumenthaler C, Wuthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol 273:283–298

Herrmann T, Guntert P, Wüthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J Mol Biol 319:209–227

Huang Y, Powers R, Montelione G (2005) Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. J Am Chem Soc 127:1665–1674

Huang YJ, Tejero R, Powers R (2006) A topology-constrained distance network algorithm for protein structure determination from NOESY data. Proteins Struct 62:587–603

Lange OF (2013) Implementation of automatic NOE assignment in Rosetta. J Biomol NMR. doi:10.1007/s10858-014-9833-3

Lange OF et al (2012) Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. Proc Natl Acad Sci USA 109:10873–10878

Laskowski RA et al (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8:477–486

Linge JP, Habeck M, Rieping W, Nilges M (2003a) ARIA: automated NOE assignment and NMR structure calculation. Bioinformatics 19:315–316

Linge JP et al (2003b) Refinement of protein structures in explicit solvent. Proteins 50:496–506

Mao B, Guan R, Montelione GT (2011) Improved technologies now routinely provide protein NMR structures useful for molecular replacement. Structure 19:757–766

Moseley HN, Montelione GT (1999) Automated analysis of NMR assignments and structures for proteins. Curr Opin Struct Biol 9:635–642

Nilges M, Macias MJ, Odonoghue SI, Oschkinat H (1997) Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from beta-spectrin. J Mol Biol 269:408–422

Raman S et al (2010) Accurate Automated Protein NMR Structure Determination Using Unassigned NOESY Data. J Am Chem Soc 132:202–207

Rosato A et al (2009) CASD-NMR: critical assessment of automated structure determination by NMR. Nat Methods 6:625–626

Rosato A et al (2012) Blind testing of routine, fully automated determination of protein structures from NMR data. Structure 20:227–236

Schmidt E, Guntert P (2012) A new algorithm for reliable and general NMR resonance assignment. J Am Chem Soc 134:12817–12829

Serrano P et al (2012) The J-UNIO protocol for automated protein structure determination by NMR in solution. J Biomol NMR 53:341–354

Vernon R, Shen Y, Baker D, Lange OF (2013) Improved chemical shift based fragment selection for CS-Rosetta using Rosetta3 fragment picker. J Biomol NMR 57:117–127

Vriend G (1990) WHAT IF: a molecular modeling and drug design program. J Mol Graph 8:52–56

Wüthrich K (1986) NMR of proteins and nucleic acids. The George Fisher Baker non-resident lectureship in chemistry at Cornell University. Wiley, p 292

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-017-02592-z OPEN

Automated NMR resonance assignments and structure determination using a minimal set of 4D spectra

Thomas Evangelidis[1], Santrupti Nerli[2,3], Jiří Nováček[1], Andrew E. Brereton[4], P. Andrew Karplus[4], Rochelle R. Dotas[5], Vincenzo Venditti[5,6], Nikolaos G. Sgourakis[3] & Konstantinos Tripsianes[1]

Automated methods for NMR structure determination of proteins are continuously becoming more robust. However, current methods addressing larger, more complex targets rely on analyzing 6–10 complementary spectra, suggesting the need for alternative approaches. Here, we describe 4D-CHAINS/autoNOE-Rosetta, a complete pipeline for NOE-driven structure determination of medium- to larger-sized proteins. The 4D-CHAINS algorithm analyzes two 4D spectra recorded using a single, fully protonated protein sample in an iterative ansatz where common NOEs between different spin systems supplement conventional through-bond connectivities to establish assignments of sidechain and backbone resonances at high levels of completeness and with a minimum error rate. The 4D-CHAINS assignments are then used to guide automated assignment of long-range NOEs and structure refinement in autoNOE-Rosetta. Our results on four targets ranging in size from 15.5 to 27.3 kDa illustrate that the structures of proteins can be determined accurately and in an unsupervised manner in a matter of days.

[1] CEITEC—Central European Institute of Technology, Masaryk University, Kamenice 5, Brno 62500, Czech Republic. [2] Department of Chemistry and Biochemistry, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [3] Department of Computer Science, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [4] Department of Biochemistry and Biophysics, 2011 Ag & Life Sciences Bldg, Oregon State University, Corvallis, OR 97331, USA. [5] Department of Chemistry, Iowa State University, 2438 Pammel Drive, Ames, IA 50011, USA. [6] Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA 50011, USA. Correspondence and requests for materials should be addressed to N.G.S. (email: nsgourak@ucsc.edu) or to K.T. (email: kostas.tripsianes@ceitec.muni.cz)

Nuclear magnetic resonance (NMR) structure determination relies on recording a network of nuclear Overhauser enhancement (NOE) restraints from multidimensional spectra[1]. Obtaining near-unambiguous assignments of long-range NOEs is challenging due to substantial overlap in the spectra which becomes more pronounced for larger proteins. This is typically addressed through first establishing the chemical shift assignments of backbone and sidechain atoms using multiple (6–10) triple-resonance spectra[2, 3], which are then used as anchors to guide the assignment of NOEs during iterative structure refinement[4]. State-of-the-art tools such as FLYA[5], PINE[6] and UNIO[7] can automate the resonance assignment and structure determination process. In principle, recording a smaller number of higher dimensionality spectra can provide a complementary approach to increase signal dispersion and resolve ambiguities[8]. With the emergence of non-uniform sampling and reconstruction methods, such datasets can be recorded in reasonable time[9]. Recent approaches for automated resonance assignments based on three- and four-dimensional (3D and 4D) NOE data make use of a known structure to guide the assignment process[10, 11]. However, for de novo structure determination, further development is needed to perform resonance assignments at the high levels of completeness and correctness that are required for NOE data-driven structure determination.

Recent methods allow for structure modeling guided by NMR chemical shifts, used as a means to optimize a physically realistic energy function that reproduces the native features of protein structures[12, 13]. Chemical shift Rosetta (CS-Rosetta) relies on backbone assignments along with Rosetta's Monte Carlo fragment assembly protocol to model protein structures in the 10–12 kDa range[12]. CS-Rosetta was superseded by resolution-adapted structural recombination (RASREC)-Rosetta, extending the size limit to 25 kDa using backbone residual dipolar couplings (RDCs) and amide NOEs[14], or to 40 kDa using sparse NOE data acquired on methyl-labeled, perdeuterated NMR samples, assigned manually[15]. In addition, the use of evolutionary information in conjunction with NMR chemical shift data can be used to model protein targets in the 25–40 kDa range[16–18]. Finally, autoNOE-Rosetta performs automated assignment of long-range NOEs and structure refinement using iterations of parallel RASREC-Rosetta calculations[19]. In all these methods, the use of advanced conformational sampling methodologies enables protein structure determination using a sparse network of restraints[20]. However, a significant bottleneck remains in establishing correct sidechain assignments at sufficient completeness levels to drive the automated assignment of long-range NOEs[20]. Moreover, the use of methyl-labeled samples requires extensive deuteration, which can be challenging for several biologically important systems[21].

Here we combine the powerful autoNOE-Rosetta approach with a new automated assignment algorithm (4D-CHAINS) in a complete pipeline for NMR structure determination. First, 4D-CHAINS utilizes two complementary experimental datasets, a 4D-TOCSY (Total Correlated Spectroscopy) and a 4D-NOESY (Nuclear Overhauser Effect Spectroscopy), to obtain near-complete resonance assignments of backbone and sidechain ^{1}H, ^{13}C and ^{15}N atoms. The resonance lists provided by 4D-CHAINS form the basis for iterative assignment of long-range NOEs and structure determination using autoNOE-Rosetta, which exploits through-space correlations recorded in two 4D-NOESY datasets, one amide to aliphatic, and one aliphatic to aliphatic. The combined approach allows us to obtain structural ensembles for proteins up to 27 kDa, without the need for deuteration or selective labeling, by leveraging the well-resolved spectral features of the 4D datasets together with Rosetta's energy function. Our NMR data and detailed analysis, performed for one benchmark case with known X-ray structure and three additional blind targets, illustrate that the new approach can consistently deliver high-resolution structural ensembles of biologically relevant proteins by greatly reducing the number of required experiments and human time spent.

Results

Development of the 4D-CHAINS assignment algorithm. Towards developing 4D-CHAINS, we recorded for four different protein targets of size from 15.5 to 27.3 kDa, a 4D HC(CC-TOCSY(CO))NH, and a 4D ^{13}C,^{15}N edited HMQC-NOESY-HSQC (HCNH) experiment. The largest protein target of size 27.3 kDa was chosen based on its apparent correlation time of ~15 ns that still allows for TOCSY transfer to occur (Supplementary Figure 1). We also recorded a 4D ^{13}C,^{13}C edited HMQC-NOESY-HSQC (HCCH) experiment to further assist in structure determination. To address the assignment problem, 4D-CHAINS uses 2D probability density maps of correlated ^{13}C–^{1}H chemical shifts to effectively identify possible spin systems (Fig. 1, Supplementary Figure 2). In particular, 4D-CHAINS combines sequential information present in the 4D-HCNH TOCSY and intraresidue information present in 4D-HCNH NOESY ^{13}C–^{1}H planes, respectively, by clustering TOCSY or NOESY peaks to Amino Acid Index Groups (AAIGs) via their common ^{15}N–^{1}H frequency (Supplementary Figure 3). 4D-CHAINS computes probability scores at several steps (amino acid-type prediction, sequential AAIG relations based on TOCSY–NOESY connectivities, alignment of peptides to the protein sequence) to yield a confidence score for a given AAIG being assigned to a specific protein residue. Finally, 4D-CHAINS uses an Overlap Layout Consensus (OLC) assembly approach adopted from genome assembly[22] to match continuous AAIG segments along the protein sequence (Fig. 2). The final assignment solutions are consistent with both the joined probability score and the OLC model.

A uniform 4D-CHAINS protocol was applied to all four targets (Fig. 3a). The algorithm mapped correctly all AAIGs to the respective protein sequences with >95% completeness (Supplementary Figure 4). The TOCSY-based assignments alone covered approximately 80% of all aliphatic chemical shifts with an error rate of <0.5% (Fig. 3b, Supplementary Figure 5). To increase the overall assignment completeness, we obtained additional information from the HCNH NOESY spectrum by employing the concept of common NOEs between successive residues[21]. The TOCSY–NOESY combination enabled more complete assignments with 94% correct aliphatic chemical shifts and a combined error rate of 1.3% (Fig. 3b, Supplementary Figures 5–7, Supplementary Table 1). The concept of common NOEs in obtaining assignments was further tested by providing 4D-CHAINS fixed ^{15}N–^{1}H assignments and the HCNH NOESY spectrum alone (Supplementary Figure 8). This assignment scenario (NOESY) allows users to extend existing backbone assignments, obtained conventionally, to cover sidechain atoms with 86% accuracy and an error rate of approximately 5% (Fig. 3b, Supplementary Figure 5).

Performance of 4D-CHAINS relative to existing methods. To test the performance of 4D-CHAINS relative to existing assignment programs, we performed calculations using a popular assignment method, FLYA[5], for all protein targets used in the current study. While 4D-CHAINS relies exclusively on the combination of 4D-HCNH TOCSY and 4D-HCNH NOESY, the FLYA algorithm is designed to combine peak patterns from any number of input spectra. Therefore, we provided FLYA with all available spectra (4D-HCNH TOCSY, 4D-HCNH NOESY, 4D-HCCH NOESY). Notwithstanding, 4D-CHAINS outperforms FLYA consistently for all four protein targets in our benchmark.

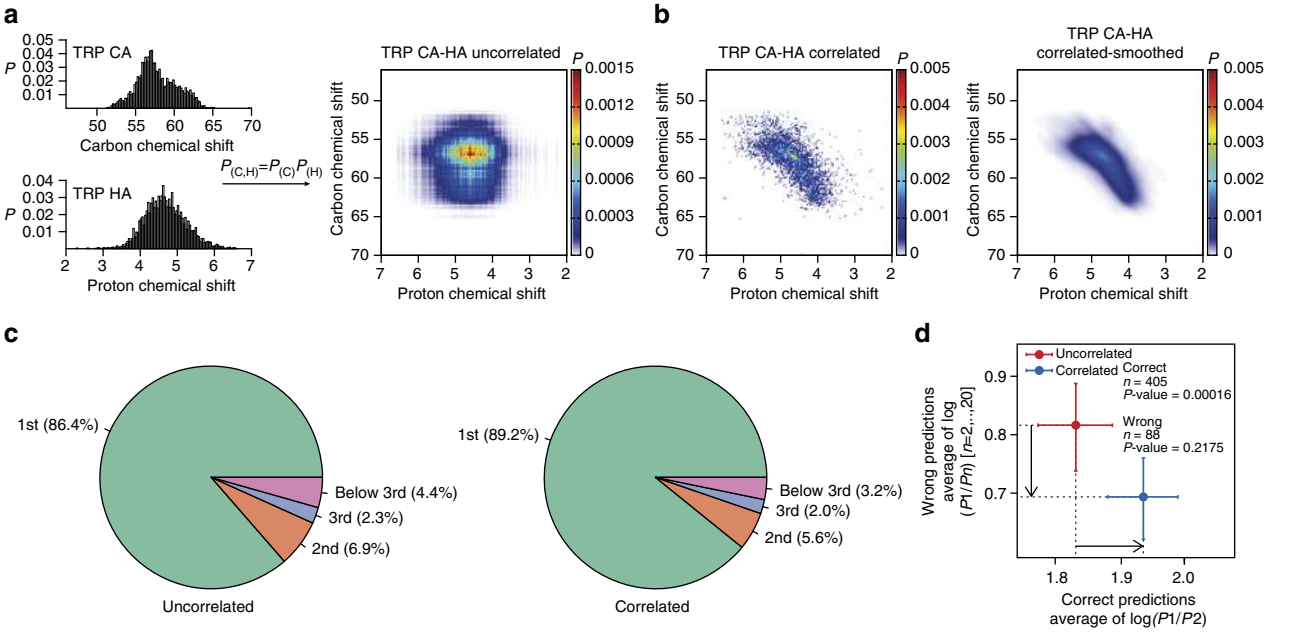


Fig. 1 Comparison of uncorrelated and correlated chemical shifts probabilities and their power in amino acid-type prediction. **a** The 1D probability distributions for the C_α and H_α atoms of Trp (left) and their joint probability distribution (right). **b** The 2D probability distribution of correlated C_α-H_α chemical shifts of Trp (left) and the corresponding smoothed 2D probability density map (right) after applying a Gaussian kernel function. Graphs were generated from the VASCO database with a bin size of 0.04 p.p.m. for protons and 0.2 p.p.m. for carbons. The color gradient scaling differs for (**a)** uncorrelated and (**b)** correlated distributions. **c** Pie charts displaying the ranking of amino acid-type predictions in the current dataset (four proteins) using uncorrelated (left) or correlated distributions (right) of ^{13}C-^{1}H chemical shifts. **d** Analysis of common (correct or wrong) predictions made by using uncorrelated (red) or correlated distributions (blue) of ^{13}C-^{1}H chemical shifts. Horizontal axis plots the mean value and standard error of correct predictions (*P*1) to the second probable (*P*2), and vertical axis plots the mean value and standard error of wrong predictions (*Pn*) to the first probable (*P*1). Pairwise *t*-test shows that the improvement in predictions made by using correlated instead of uncorrelated ^{13}C-^{1}H chemical shifts (arrows) is statistically significant for the correct ones (*P*-value = 0.00016) but not for the wrong ones (*P*-value = 0.2175). The discrepancy is attributed to the relatively small number of wrong predictions available for the test (405 vs 88)

For three proteins, namely RTT, ms6282 and Enzyme I (nEIt), FLYA outputs 90% correct assignments with 7-8% error rate, while for α-lytic protease (aLP) the number of correct assignments is limited to 25% (Supplementary Figure 5). Finally, we manually inspected and extended the 4D-CHAINS results to establish the maximum number of highly accurate assignments for all ^{13}C–^{1}H correlations that can be observed in our 4D spectra (>98%), as a "best-effort" resonance list requiring a modest time investment by a trained user. These supervised assignment lists also contain aromatic and sidechain amide chemical shifts, not considered by the automated 4D-CHAINS protocol (Supplementary Figure 6).

Iterative structure calculations using autoNOE-Rosetta. We evaluated the performance of assignments obtained using 4D-CHAINS in driving Rosetta structure determination of a 20 kDa target, aLP, for which several known X-ray structures are available in the Protein Data Bank[23]. Inspection of the X-ray structural ensemble shows a highly complex all-β fold, with two subdomains each containing a 6-stranded antiparallel β-sheet. In order to establish a "best effort" limit of the Rosetta automated NOE assignment and structure determination protocol, we first performed autoNOE-Rosetta calculations[19] using the supervised assignments together with both NOE datasets (HCNH+HCCH). Additionally, we carried out automated 4D-CHAINS/autoNOE-Rosetta structure calculations (Fig. 3a, Supplementary Figure 9) under four different scenarios, as described above (TOCSY–NOESY or NOESY assignments, each using HCNH alone or HCNH+HCCH NOEs). To evaluate the quality of the resulting structural ensembles, we used the following criteria: (i) fraction of residues converged within 2.0 Å backbone heavy atom root-mean-square deviation (RMSD) in the final ensemble, (ii) average Rosetta all-atom energies and (iii) RMSD to X-ray structure. Since the Rosetta energy function[24] has a global minimum at the native fold, lower-energy models should also exhibit higher convergence towards the native structure. We observed a good correlation between Rosetta all-atom energy, degree of convergence and structural accuracy (correlation coefficient of 0.93, Fig. 3c). Specifically, using the supervised assignments and HCNH or HCNH+HCCH NOEs, we obtained highly converged structural ensembles (>98%, computed over the core secondary structure regions) that are within 0.7 Å RMSD from the X-ray structure.

Notably, ensembles calculated using the 4D-CHAINS (TOCSY–NOESY or NOESY) automated assignments using both NOE datasets also achieved a high level of convergence (>90%), to within 1.3 and 1.7 Å RMSD from the X-ray, respectively. Using the same 4D-CHAINS assignment lists and the HCNH NOEs alone, the accuracy relative to the X-ray was slightly reduced to 1.7 and 1.9 Å, respectively, while the convergence decreased to approximately 86%, but the models still recapitulated the protein fold and β-sheet topology. This trend is highlighted in a superposition of the lowest-energy aLP model sampled in each calculation on the X-ray structure reference (Fig. 3d).

Convergence of aLP structures towards the X-ray reference. To evaluate the relative accuracy and precision of NOE-driven structure determination approaches using different input

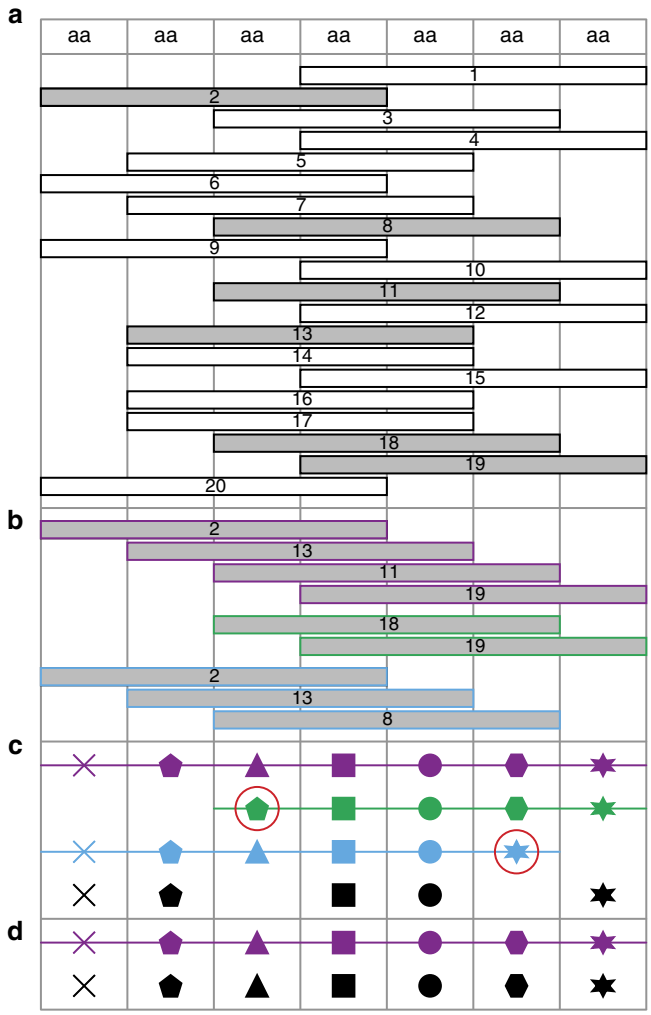


Fig. 2 Schematic representation of 4D-CHAINS decision making using Overlap Layout Consensus assembly. **a** Chains of AAIGs of given length *L* are translated to peptides and aligned to the query sequence. **b** Contigs of overlapping chains (overlap length *L*-1) are generated. Peptides that do not form contigs are discarded. **c** Contigs are lined up and absolute consensus AAIGs (shapes) are identified. If there is agreement with the associated confidence score (for details see Methods), then AAIGs are assigned to the corresponding amino acids of the sequence (black shapes). **d** Assigned AAIGs are restrained in next rounds of decision making. False peptides and contigs do not longer form and using the same rules additional AAIGs are assigned (triangle, polygon)

assignments and NOE datasets (HCNH or HCNH+HCCH), we performed a single joint Ensemblator[25, 26] analysis of the 12 resulting aLP structural ensembles (six each using Rosetta or CYANA[27]) and a set of 51 X-ray structures from the CoDNaS[28] database (see Methods). A dimensionality-reduced visualization of the relationships between the models (Fig. 4a) reveals that the Rosetta models are consistently closer to the X-ray models than the corresponding CYANA models generated using the same datasets. Overall, the Rosetta models show better convergence, and convergence for all groups correlates strongly with their similarity to the X-ray structures (Fig. 4b). The two Rosetta-generated ensembles based on supervised assignments are nearly equivalent and show the best convergence and greatest similarity to the X-ray structures (Fig. 4a, b). Here, the fragment-based structure refinement in Rosetta allows the generation of highly accurate ensembles from HCNH NOEs alone, which is not feasible using standard simulated annealing in CYANA. In particular, poor convergence and low similarity to the X-ray ensemble are seen for models calculated by CYANA from the HCNH NOEs alone (Fig. 4a, b; orange and red circles). Conversely, the Rosetta ensembles produced from a single NOESY dataset (HCNH) are in good agreement with the X-ray ensemble (Fig. 4c), and quantitative comparison shows that the structural variability pattern along the protein chain is rather similar, although the NMR ensemble typically has a greater variability than the X-ray ensemble (Fig. 4d). These results suggest that the uncertainty of atom positions in solution correlates with variability associated with different crystal packing environments.

Consistent blind structure determination of protein targets. To further test our method in a fully unbiased manner we performed blind structure calculations for three additional protein targets, RTT[29, 30], ms6282 and nEIt of sizes 133, 145 and 248 amino acids (aa), respectively (Table 1). To establish a baseline performance, we carried out CS-Rosetta calculations guided by chemical shifts alone[15], as well as reference CYANA calculations using both input NOE datasets (HCNH+HCCH). With the exception of the smallest target (RTT), the resulting CS-Rosetta models failed to converge (Supplementary Figure 10) and instead sampled conformations with sub-optimal energies (Fig. 5; right column, black). Conformational sampling is drastically improved in autoNOE-Rosetta calculations guided by both supervised or automated 4D-CHAINS assignments, and the resulting structural ensembles are very similar for all targets (Fig. 5; left column). For the largest target, the 27.3 kDa Enzyme I from *Thermoanaerobacter tengcongensis*, NOE contacts provided sufficient constraints to elucidate the structure of the individual domains, but the overall orientation of the two domains was not converged due to the lack of contacts at their interface (domain A, defined by residues 1–143 and domain B, defined by residues 144–248) (Supplementary Figure 11). Here, the use of ^{15}N–^{1}H residual dipolar couplings allowed us to sample further lower energies, and obtain better convergence by restraining the relative orientation of the two domains (Fig. 5d, Supplementary Figure 11d).

Towards evaluating the effect of different levels of assignment completeness on the performance of autoNOE-Rosetta, we carried out benchmark calculations by randomly removing entries from our "best effort" supervised assignment lists for target aLP and found that autoNOE-Rosetta can identify correct protein fold from as low as 60–70% sidechain assignments. In addition, we performed a detailed comparison of assigned NOE contacts and Rosetta energy distributions, relative to control calculations guided by the supervised assignments. We observe that the use of fully automated assignments results in a small decrease in the total number of NOE contacts identified by Rosetta (approximately 80% for all targets). Furthermore, we obtain similar distributions of assigned NOE contacts among residue pairs in the protein sequence (Fig. 5; middle column). The respective lowest-energy models are built using hundreds of automatically assigned, long-range NOE restraints and exhibit a minimal number of violations (1–4%) involving pairs of atoms that are typically within 1 Å from their estimated upper distance limits (Supplementary Table 2). Given that methyl–methyl NOE contacts play a critical role in defining the hydrophobic core of the protein, we found that ~25% of the total contacts identified by autoNOE-Rosetta are contributed by methyl NOEs for structure calculations using supervised or automated 4D-CHAINS assignments (Supplementary Table 3). Finally, the distributions of energies among the 100 best sampled structures are generally shifted relative to RASREC-Rosetta and show good overlap with their supervised counterparts (Fig. 5; right column).

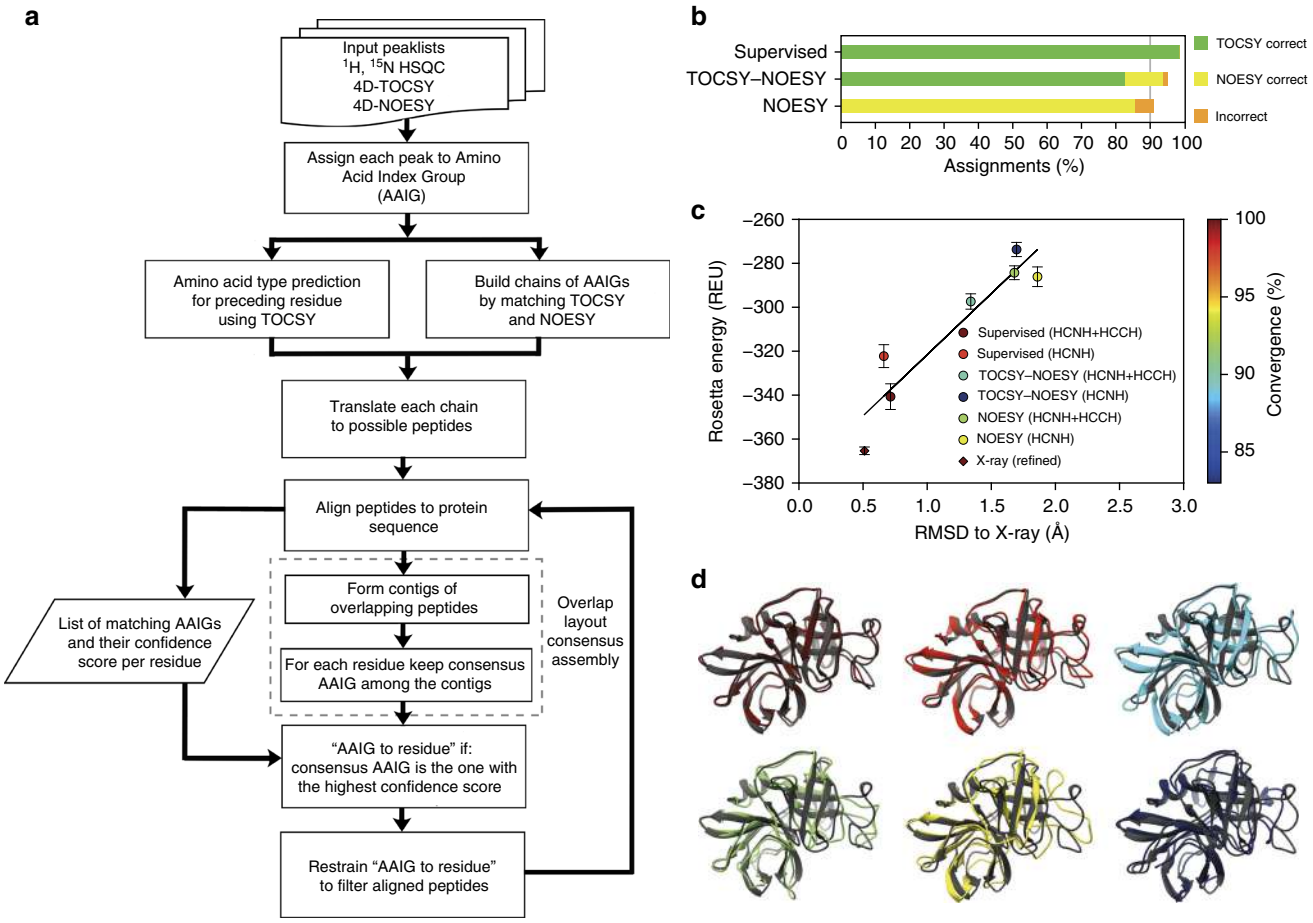


Fig. 3 Automated structure determination using 4D-CHAINS/autoNOE-Rosetta. **a** Flowchart of the 4D-CHAINS algorithm for automated NMR resonance assignment from two 4D spectra (TOCSY and NOESY). **b** Quality of 4D-CHAINS assignments for supervised, TOCSY–NOESY, and NOESY settings, expressed as the average for the four different protein targets. **c**, **d** Performance of different 4D-CHAINS assignment scenarios for a 198 aa protein, α-lytic protease, calculated using autoNOE-Rosetta. **c** Goodness of structural ensembles is measured using the Rosetta all-atom energy function, backbone heavy atom RMSD to X-ray structure (PDB ID 1P01) and degree of structural convergence. Average energy values (in Rosetta Energy Units (REU)) for ensembles calculated using indicated data/assignment scenarios, with errors bars shown at 1 standard deviation. Also shown is the average energy of 10 locally refined X-ray structures (diamond)—refinement adapts the X-ray structure to the local optimum of the Rosetta energy, with a minimum change in RMSD (0.5 Å). The color of points represents the average % of converged residues in each ensemble, according to the color scale on the right. **d** Lowest-energy structures in each ensemble (shown in the same color as the points in **c**) superimposed on the X-ray reference structure (gray). Images of structures were produced using Chimera (https://www.cgl.ucsf.edu/chimera)

Discussion

NMR remains the only biophysical technique that can deliver high-resolution structures of proteins and other biomolecules in their functional, aqueous environment, which constitutes the basis for studying interactions with other molecules and therapeutic compounds. However, standard approaches for NMR resonance assignment rely on recording several complementary datasets which can be limiting for larger, more complex systems due to increased resonance overlap and require a significant time investment by a trained expert to analyze the spectra and establish a complete list of resonance assignments aided by computational tools[31]. Established methods to overcome this problem utilize selective isotopic labeling[32], which can be limiting in terms of the information content present in the NMR data, expensive and challenging to perform for certain systems.

Here, we propose an automated approach for full structure determination using 2–3 4D NMR spectra recorded on a ^{13}C, ^{15}N uniformly labeled sample. First, 4D-CHAINS addresses the assignment problem in an efficient and highly robust manner, yielding the correct assignments for at least 95% of residues and error rates of less than 1.5% (Supplementary Table 1). It is further worth noting that the vast majority of resonances corresponding to sidechain methyls, which are important probes in identifying the protein fold, are correctly assigned by our method. Therefore, the use of 4D-CHAINS allows near-complete assignment of sidechain methyls without the need for site-specific labeling on a perdeuterated background[33] (Supplementary Table 4). Second, autoNOE-Rosetta uses a highly parallelizable iterative algorithm run on a computer cluster to perform assignment of long-range NOEs alongside the structure determination process. The full pipeline takes approximately 10–12 days to execute for a typical protein sample, including the time needed for NMR data acquisition, and requires minimum supervision.

In addition to recapitulating the correct protein fold, autoNOE-Rosetta models obtained using the 4D-CHAINS assignments show accurate placement of sidechains for most residues in the protein structure (Supplementary Table 5). Specifically, close inspection of sidechain conformations in the Rosetta ensembles computed using the supervised assignments shows good overall agreement with the X-ray rotamers for most

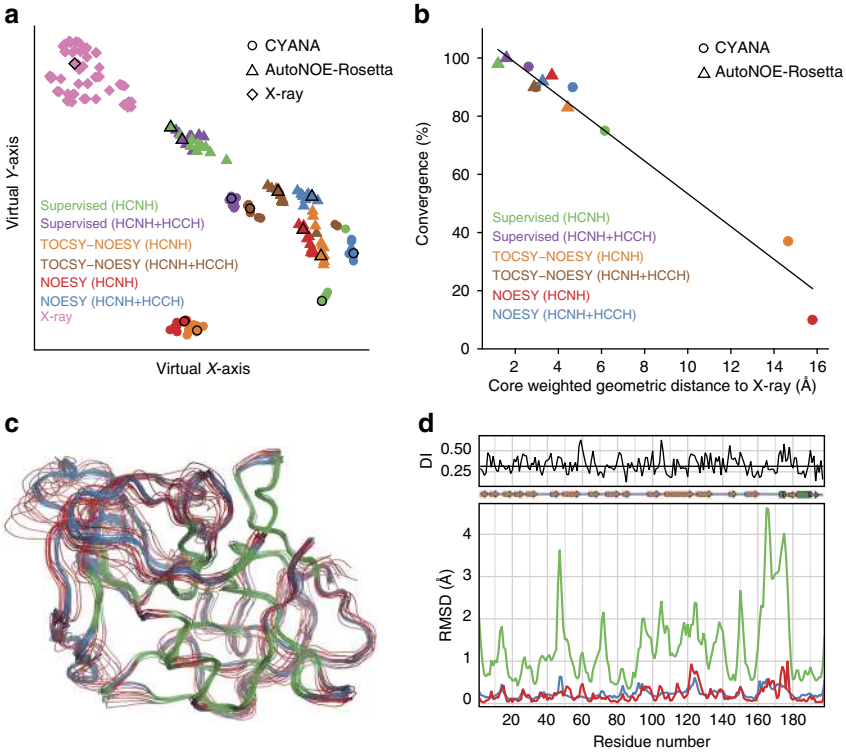


Fig. 4 Ensemblator[25, 26] analysis of α-lytic protease NMR and X-ray ensembles. **a** t-SNE dimensionality reduction results showing the relationships of the aLP models. The shape and color of each point (see in plot key) convey the method and the data used to generate that model; a black outline highlights the most representative (exemplar) for each type of model. **b** For each type of NMR ensemble, the convergence of the ensemble during refinement is compared to the distance of its exemplar to that of the X-ray ensemble. Shapes and colors of each point (see in plot key) convey the method and the data used to generate that model. The best fit to these data (black line) has an adjusted R^2 of 0.96. **c** Wire diagram showing the traces of the backbone heavy atoms for the 4D-CHAINS/autoNOE-Rosetta ensemble obtained using supervised assignments and HCNH NOEs (red) and the X-ray ensemble (blue). The defined core atoms (21.4% of all atoms using a 1.6 Å cutoff) are shown in green. **d** Shown are the per residue global backbone heavy atom variation (bottom panel) for the 4D-CHAINS/autoNOE-Rosetta ensemble obtained using supervised assignments and HCNH NOEs (green) and the X-ray ensemble (blue), along with the closest approach distance for models in each group (red). The discrimination index (top plot, black line) reveals regions of similarity (low values) and difference (high values); the median discrimination (DI) index is indicated by a horizontal line

buried residues (>10 Å^2 BSA), while using the fully automated assignments results in a small decrease (<10%) in accuracy relative to the models derived using supervised assignments (Supplementary Figure 12 and Supplementary Table 5). Finally, an analysis of long-range NOEs assigned by autoNOE-Rosetta versus predicted from the X-ray structure using a 5.5 Å distance cutoff between all pairs of protons shows good recovery of crystallographic contacts at levels of 67–86%, which are distributed across the entire protein fold (Fig. 6). Taken together, our results underpin that the automated 4D-CHAINS/autoNOE-Rosetta approach yields models that accurately capture the correct global fold as well as atomic features of the native structure.

Overall, the convergence of structures obtained using supervised assignments for the three target proteins, RTT, ms6282 and aLP, are better than or comparable to the convergence of structures obtained using 4D-CHAINS automated assignments, as expected (Table 1). Notably, for Enzyme I, autoNOE-Rosetta achieves a higher level of structural convergence using automated assignments due to enhanced resampling of the correct protein fold during the early stages of the autoNOE-Rosetta structure calculation process. Overall, our results suggest that the fully automated assignment process introduced by 4D-CHAINS has a minimal impact on the performance and quality of the derived structural ensembles by autoNOE-Rosetta, which remain highly consistent with all available input data.

Relative to CYANA, autoNOE-Rosetta can achieve a similar degree of structural convergence using the same input resonance assignments with both the aliphatic and amide NOE peak lists. Although the total number of structurally degenerate HCNH +HCCH long-range NOE contacts identified by CYANA is higher by (i) ~10% for aLP and ms6282 and (ii) ~25% for RTT and nEIt (Supplementary Figure 13), for three targets, RTT, ms6282 and aLP, the degree of structural convergence achieved by autoNOE-Rosetta is comparable to CYANA; while for the largest target, Enzyme I (nEIt), the autoNOE-Rosetta ensemble is significantly more converged towards the correct fold (Supplementary Figure 14). Generally, the structural ensembles determined using autoNOE-Rosetta are closer to the nearest PDB reference structures by approximately 0.5 Å for RTT, 0.2 Å for ms6282, 0.5 Å for aLP and >2.2 Å for nEIt relative to the structures predicted by CYANA (Supplementary Figure 14). When using the amide NOEs alone together with either automated or supervised resonance assignments, CYANA does not yield converged models, while autoNOE-Rosetta can still deliver models showing the correct protein fold, albeit with reduced convergence relative to calculations performed using both input peak lists, as shown in detail for aLP (Fig. 3) and as outlined for all other targets (Table 1).

In summary, we demonstrate that 4D-CHAINS provides highly accurate and near-complete NMR resonance assignments from two 4D spectra, which are effective in guiding high-resolution structure determination using autoNOE-Rosetta. Our results on four targets in the 15.5–27.3 kDa range indicate that the use of our automated pipeline has a minimal impact on the

Table 1 Statistics of autoNOE-Rosetta structural ensembles computed using supervised and automated 4D-CHAINS assignments

Protein	Assignment data	Peaks	RDCs used	No. of residues	Average Rosetta energy (REU)	[a]Fraction of residues converged (%)	Mean RMSD (Å) to average structure
RTT	Supervised	HCNH		133	−239.66	98	0.86
		HCNH +HCCH			−247.31	98	0.71
	TOCSY-NOESY	HCNH		133	−243.63	99	1
		HCNH +HCCH			−243.32	95	0.94
ms6282	Supervised	HCNH		145	−263.83	92	1.15
		HCNH +HCCH			−252.8	97	0.9
	TOCSY-NOESY	HCNH		145	−262.95	93	1.19
		HCNH +HCCH			−262.57	93	0.94
aLP	Supervised	HCNH		198	−324.46	99	0.84
		HCNH +HCCH			−328.5	100	0.64
	TOCSY-NOESY	HCNH		198	−274.79	86	1.49
		HCNH +HCCH			−300.55	91	1.27
nEIt	Supervised	HCNH	Yes[b]	248	−487.79	60	3.75
		HCNH +HCCH	Yes[b]		−475.61	84	1.4
	TOCSY-NOESY	HCNH	Yes[b]	248	−479.54	60	4.07
		HCNH +HCCH	Yes[b]		−491.05	91	1.36

[a] Convergence statistics calculated over core residues
[b] One RDC dataset was used to improve structure convergence
Average Rosetta energies (in Rosetta Energy Units (REU)) reported over 10 lowest-energy structures
All backbone heavy-atom RMSD relative to the average structure calculated over core residues

precision and quality of the resulting structural ensembles, while allowing for a tremendous reduction in human effort and NMR spectrometer time. Lastly, our structural evaluation criteria, in terms of convergence and Rosetta all-atom energy, can clearly distinguish the correct structures, allowing our protocol to be used extensively for generating high-quality models in a truly unsupervised manner. Therefore, our combined approach could be of great practical utility in both high-throughput structural determination projects[34] and NMR-based screening for small-molecule and protein–protein interactions[35].

Methods

NMR sample details. For each uniformly ^{13}C-, ^{15}N-labeled protein sample, the concentration, buffer composition and NMR data collection temperature are as follows:

The 0.8 mM RTT in 35 mM potassium phosphate (pH 6.8), 100 mM KCl, 5% D_2O, 25 °C.

The 1.2 mM ms6282 in 50 mM sodium phosphate (pH 6.5), 150 mM NaCl, 7% D_2O, 25 °C.

The 2.0 mM aLP in 10 mM deuterated sodium acetate (pH 4.0), 50 mM NaCl, 8% D_2O, 25 °C.

The 1.8 mM nEIt in 20 mM sodium phosphate (pH 6.5), 100 mM NaCl, 5% D_2O, 37 °C.

NMR data collection. For each protein target, a set of three sparsely sampled 4D NMR experiments was acquired on 850 or 950 MHz Bruker Avance III spectrometers equipped with ^{1}H/^{13}C/^{15}N TCI cryogenic probehead with *z*-axis gradients. All NMR spectra were recorded at CEITEC Josef Dadok National NMR Centre using pulse sequences adopted and modified from Bruker library. The 4D HC(CC-TOCSY(CO))NH experiment was acquired with chemical shift evolution performed in semi-constant time manner in t_1 ($^{1}H_{ali}$) and t_3 (^{15}N) and using FLOPSY16 spin-lock of 12 ms that yielded the best overall signal-to-noise ratio. The spectral widths were set to 12,500 (acq) × 2000 (^{15}N) × 8000 ($^{13}C_{ali}$) × 6250 ($^{1}H_{ali}$) Hz and maximal acquisition times in the indirectly detected dimensions were set to 50 ms for ^{15}N, 10 ms for $^{13}C_{ali}$ and 16 ms for $^{1}H_{ali}$. The experiment was acquired with 16 scans per increment and single-scan recycling delay of 1.0 s. The overall number of 1536 points was collected in the acquisition dimension and 2500 hypercomplex points were sparsely distributed over the indirectly detected dimensions. Prior to recording full 4D HC(CC-TOCSY(CO))NH experiment, we recorded the ^{15}N/^{1}H 2D plane of the experiment using a full (incremental) sampling list since our methodology is applicable if the number of signals observed in the 2D plane are ≥50% of expected, based on a standard, sensitivity-enhanced 2D ^{15}N/^{1}H HSQC experiment. In the 4D ^{13}C,^{15}N edited HMQC-NOESY-HSQC (HCNH) experiment, the HMQC building block was used to transfer the magnetization between ^{1}H (t_1) and ^{13}C (t_2) with evolution of the ^{1}H chemical shift in semi-constant time manner during both transfer and refocusing of magnetization. The magnetization transfer between ^{1}H (t_4) and ^{15}N (t_3) was designed using a reverse HSQC building block after the 70 ms NOESY mixing time. The data were collected with spectral widths set to 12,500 (acq) × 2000 (^{15}N) × 8000 (^{13}C) × 10,000 (^{1}H) Hz, respectively. The maximal evolution times in the indirectly detected dimensions were set to 50 ms for ^{15}N, 10 ms for ^{13}C and 20 ms for ^{1}H. The experiment was acquired using 1.0 s single-scan recycle delay and 8-step phase cycle with 8 scans per increment. In all, 1536 complex points were acquired in the direct dimension and the overall number of 5000 hypercomplex points was non-uniformly distributed over the indirectly detected dimensions. The 4D ^{13}C,^{13}C edited HMQC-NOESY-HSQC (HCCH) experiment uses the same HMQC-type building block as described for the HCNH noesy experiment (see above) within the first ^{1}H (t_1), ^{13}C (t_2) transfer of magnetization. The second ^{1}H (t_4), ^{13}C (t_3) transfer of magnetization following the 70 ms NOESY mixing time is performed using HSQC building block utilizing gradients. Data were collected with spectral widths set to 12,500 (acq) × 8000 (^{13}C) × 8000 (^{13}C) × 10,000 (^{1}H) Hz, and the maximal acquisition times in the indirectly detected dimensions were set to 20 ms for ^{1}H (t_1) and 10 ms for ^{13}C (t_2, t_3). The experiment was acquired with 8 repetitions per increment and 1.0 s single-scan recycling delay. The overall number of 1536 complex points was collected in the acquisition dimension and 5000 hypercomplex points were distributed over the indirectly detected dimensions. For each spectrum the NMR acquisition time was 4 days. From our setup, we can observe that the experimental time needed to acquire three 4D non-uniform sampling spectra is comparable to the total acquisition time of several conventional 3D experiments. However, the analysis of 3D experiments is laborious and further complicated by resonance overlap, which becomes more pronounced with increasing target size. Thus, from the user's standpoint, it is preferable to operate using a pair of complementary experiments which yield the same information in a higher-dimensionality dataset. Finally, our benchmark data illustrate that any additional relaxation losses during the extra chemical shift evolution step needed to acquire the fourth indirect dimension are not prohibitive for highly concentrated samples of stable proteins, which can still yield very rich datasets. All the pulse sequences used in our experiments are available upon request.

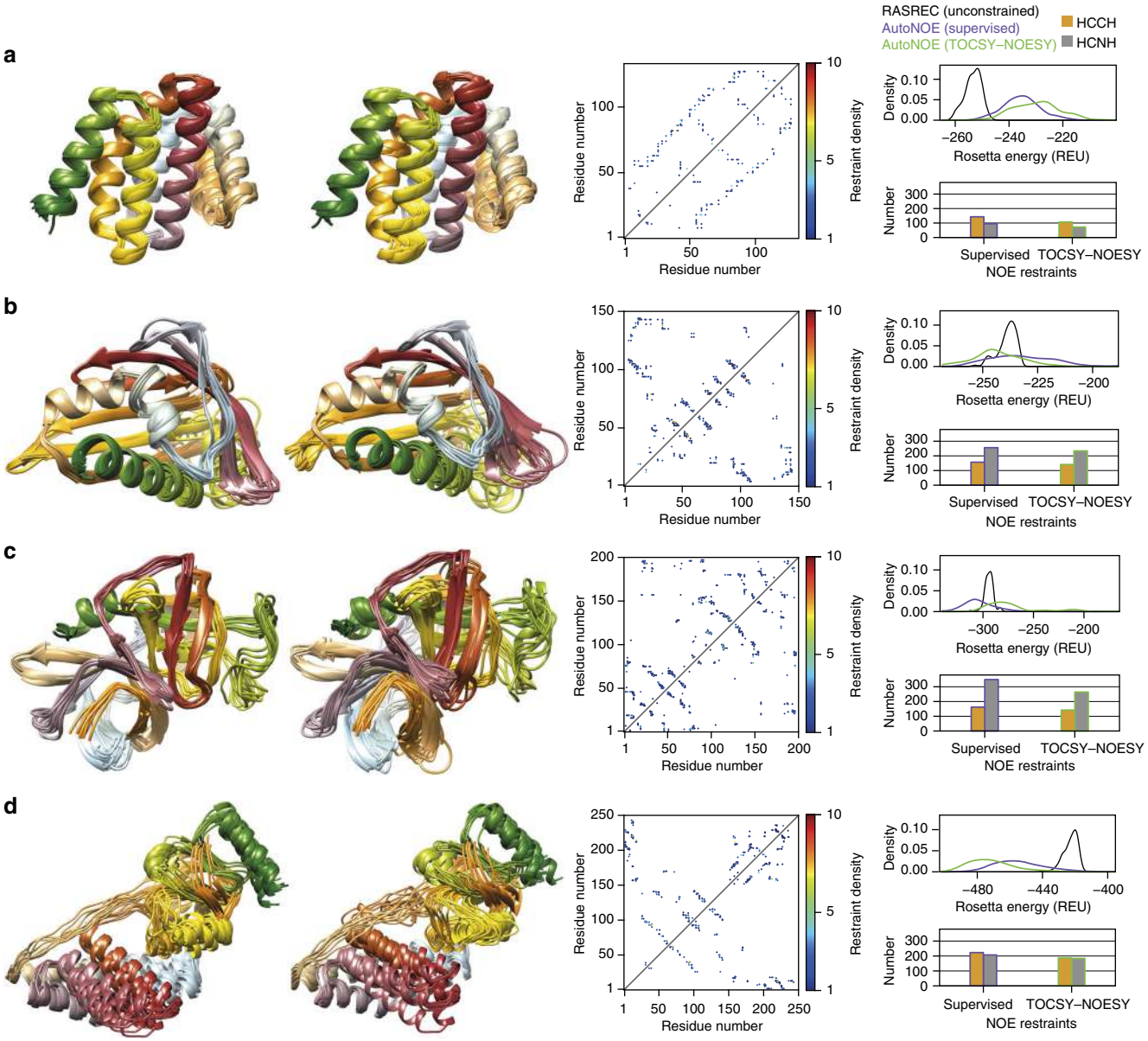


Fig. 5 Comparison of structural ensembles calculated from supervised versus fully automated assignments. **a** Rtt103 (RTT, 133 aa), **b** KanY (ms6282, 145 aa), **c** α-lytic protease (aLP, 198 aa) and **d** Enzyme I (nEIt, 248 aa). Columns 1 and 2: autoNOE-Rosetta ensembles of 10 lowest-energy structures guided by "best effort" supervised assignments or by automated 4D-CHAINS assignments (TOCSY–NOESY), respectively. Column 3: Sequence map of distance restraints assigned by autoNOE-Rosetta in iterative structure refinement calculations. Here, the upper triangular region shows restraints obtained using supervised assignments, while the lower triangular region using automated 4D-CHAINS assignments. Column 4: Rosetta energy (in Rosetta Energy Units (REU)) distributions and total numbers of assigned long-range restraints. The energy distribution was computed from the 100 lowest-energy structures sampled during the final stage of autoNOE-Rosetta calculations using supervised assignments (purple), 4D-CHAINS assignments (green) and chemical shift fragment-based RASREC-Rosetta calculations without NOEs (black). The bars represent the total number of HCNH (amide to aliphatic) and HCCH (aliphatic to aliphatic) long-range NOE restraints assigned by autoNOE-Rosetta, including ambiguous restraints derived for different stereo-specific groups. RDCs were used to obtain converged Enzyme I structures with respect to the orientation of the two domains reported in row **d**. Images of structural ensembles were produced using Chimera (https://www.cgl.ucsf.edu/chimera)

Sparse sampling and data processing. The on-grid Poisson disc sampling[36] was utilized in the present application to distribute individual acquisition points in the indirectly detected dimensions. This sampling scheme introduces distances between the generated time points and has been shown to reduce the level of sampling artifacts in the direct vicinity of signal after the reconstruction[36].

The 4D data were processed using sparse Fourier transform algorithm[37] to check the data quality. Final processing was performed in an iterative manner using the Signal separation analysis approach as implemented in the program cleaner4d[37] (SSA package). Prior to processing with the cleaner4d program, the data were square cosine weighted in the directly acquired dimension and zero-filled to 2 k points using NMRPipe/NMRDraw 3.0[38]. The 4D spectra were analyzed in Sparky[39].

Peak picking. Peaks were picked automatically and curated manually using a restricted peak picking strategy. First, the 4D-HCNH NOESY spectrum was picked at a user-defined noise level using both ^{15}N,^{1}H- and ^{13}C,^{1}H-HSQC peaks as filters. Then the 4D-HCNH TOCSY spectrum was picked using the 4D-NOESY peaks as filters. Accordingly, all planes were inspected simultaneously in all spectra and picked artifacts were removed. Synchronization of all four shared dimensions in the spectra allows for a highly efficient peak picking and curation process.

Measurement of RDC restraints for nEIt. Backbone amide $^{1}D_{NH}$ RDCs were measured by taking the difference in $^{1}J_{NH}$ scalar couplings in aligned and isotropic media[40]. The alignment media employed was phage pf1 (16 mg ml^{-1}; ASLA

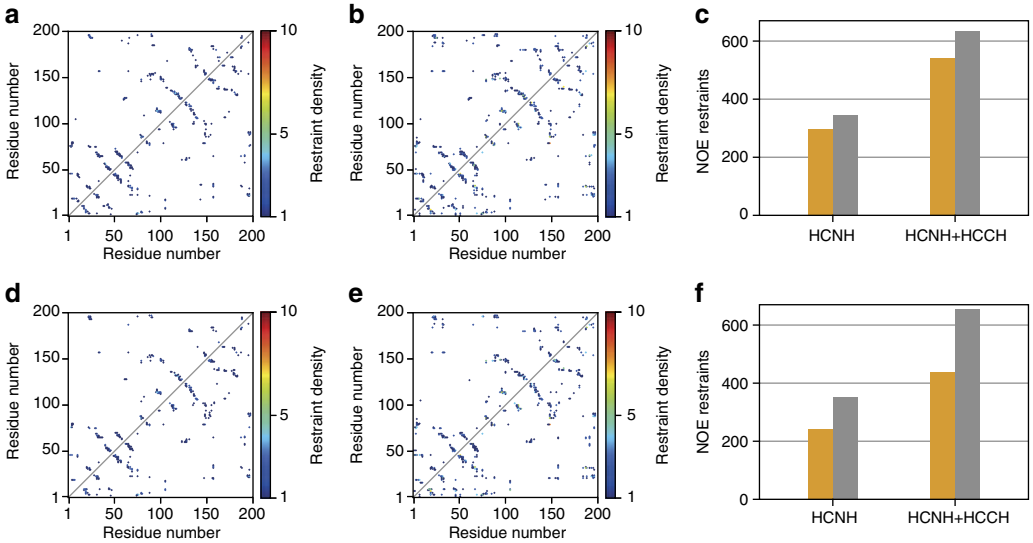


Fig. 6 Comparison of assigned NOE contacts versus predicted from X-ray structure. Contacts shown as a function of residue pairs along the sequence of α-lytic protease. Upper triangular region shows NOE contacts identified during iterative structure refinement by autoNOE-Rosetta, while that of lower triangular region represents expected contacts as predicted from the X-ray structure (PDB ID 1P01) using a 5.5 Å distance cutoff between all possible proton atom pairs and further removing redundancies due to chemically equivalent protons. Different combinations of input assignments and NOE datasets used are shown as follows. **a** Supervised assignments with HCNH NOEs. **b** Supervised assignments with HCNH+HCCH NOEs. **c** Total number of NOE restraints assigned in **a** and **b** (orange) versus predicted from X-ray structure (gray). **d** 4D-CHAINS TOCSY-NOESY automated assignments with HCNH NOEs. **e** 4D-CHAINS TOCSY-NOESY automated assignments with HCNH+HCCH NOEs. **f** Total number of NOE restraints assigned in **d** and **e** (orange) versus predicted from X-ray structure (gray)

Biotech)[41], and $^{1}J_{NH}$ couplings were measured using the ARTSY pulse scheme[42]. NMR measurements were performed on a Bruker 800 MHz spectrometer equipped with a *z*-shielded gradient triple-resonance cryoprobe. Spectra were processed using NMRPipe[38] and analyzed using the program Sparky[39].

Automated resonance assignment using 4D-CHAINS. 4D-CHAINS is an automated resonance assignment algorithm for backbone and sidechain chemical shifts of proteins. As input it requires the protein sequence in fasta format and peak lists from ^{1}H,^{15}N HSQC (root), 4D HC(CC-TOCSY(CO))NH and 4D ^{13}C,^{15}N edited HMQC-NOESY-HSQC experiments in sparky format. 4D-CHAINS algorithm tackles the assignment problem in a conventional way[43, 44]. The distinct feature of 4D-CHAINS is that all available aliphatic ^{13}C–^{1}H coupled frequencies are used for amino acid-type prediction and sequential connectivities, drastically decreasing the ambiguity level (Supplementary Figure 9).

4D-CHAINS overall assignment accuracy depends on chemical shift statistics, currently available in the form of one-dimensional (1D) distributions of proton or carbon resonances for every atom of the 20 amino acids (Fig. 1a). Since the 4D spectra provide direct information on ^{13}C–^{1}H correlated chemical shifts, we reasoned that statistical correlated chemical shift distributions would improve 4D-CHAINS performance in addressing the assignment problem. For the different ^{13}C–^{1}H moieties of every amino acid, correlated chemical shifts maps were generated from the VASCO-corrected data[45]. VASCO dataset was chosen instead of the larger BMRB dataset, because chemical shift values of aliphatic carbons that were improperly referenced have been corrected, thus avoiding distortion of the information content used. The resulting 2D probability distributions have bins with zero frequency, due to the relatively small sample size (Fig. 1b). Therefore, we created probability density maps by applying a Gaussian kernel function, given by Eq. 1, to estimate the density at any point

$$G_{(H_0,C_0)} = \frac{1}{2\pi n h_H h_C} \sum_{i=1}^{n} e^{-\frac{1}{2}\left[\frac{(H_i-H_0)^2}{h_H^2}+\frac{(C_i-C_0)^2}{h_C^2}\right]} \quad (1)$$

where n is the data size, h_H and h_C the bandwidth for the proton and the carbon dimension, respectively. For optimal bandwidth selection we used Scott's rule of thumb $h = n^{-1/6}$. Based on our analysis, the 2D probability density maps of correlated chemical shifts provide improved predictive power when compared to joint probabilities derived from 1D histograms of proton and carbon chemical shifts (Fig. 1c, d).

4D-CHAINS is written in Python programming language and consists of two modules: NH-mapping module and atom-type assignment module. As output, it provides TOCSY and NOESY (intraresidue and sequential) assignments of the input 4D peak lists allowing visual verification of results, and a chemical shift list in XEASY format that can be input together with NOESY peak lists to automated structure determination software.

At first, 4D-CHAINS clusters the 4D-HCNH TOCSY and 4D-HCNH NOESY peaks via the common root resonance (^{15}N–^{1}H) they share to generate AAIGs of ^{13}C–^{1}H correlated chemical shifts. For a given root resonance, the TOCSY AAIG provides sequential information, that is, the ^{13}C–^{1}H aliphatic resonances of the previous amino acid in the sequence (*i*-1), whereas the NOESY AAIG reports on any ^{13}C–^{1}H moiety that is in close spatial proximity. By virtue of NOE distance dependence, the NOESY AAIG contains most, if not all, of the intraresidue ^{13}C–^{1}H resonances (*i*).

For each TOCSY AAIG, 4D-CHAINS calculates the probability of an amino acid type for the preceding residue in the protein sequence using a probabilistic model[46]. Let us denote any amino acid of the 20 types by AA and the set of correlated chemical shifts in a TOCSY AAIG by CCS. The conditional probability *P* (AA|CCS) to get an amino acid type given the observed C–H resonances is highlighted in Eq. 2

$$P(\mathrm{AA}|\mathrm{CCS}) = \frac{P(\mathrm{CCS}|\mathrm{AA})P(\mathrm{AA})}{P(\mathrm{CCS})} \quad (2)$$

where *P*(CCS|AA) is the conditional probability of ^{13}C–^{1}H resonances for a given amino acid type, *P*(AA) is the prior probability of finding the given amino acid type in the protein sequence independent of the observed ^{13}C–^{1}H resonances, and *P* (CCS) is the sum of the *P*(CCS|AA) terms over the 20 amino acid types. *P*(CCS|AA) can be accurately estimated for any amino acid type using the probability density maps of ^{13}C–^{1}H correlated chemical shifts (Supplementary Figure 2). For a given number of ^{13}C–^{1}H frequencies in a TOCSY AAIG all permutations of atom-type combinations are considered in calculating the probability for amino acids with possible atom types equal or larger to the TOCSY frequencies. In practice, however, only a small number of combinations is computed, because many ^{13}C–^{1}H frequencies have non-zero probability only for distinct atom types of any amino acid (Supplementary Figure 2). *P*(CCS|AA) is considered the most probable combination, expressed as the product of probabilities of each ^{13}C–^{1}H pair belonging to different atom types of a given amino acid. If the number of TOCSY ^{13}C–^{1}H pairs is larger than the expected atom types of a given amino acid then *P* (CCS|AA) is set to zero. For every TOCSY AAIG several amino acid-type predictions are made and ranked according to their conditional probabilities. Depending on the TOCSY transfer efficiency, amino acids with long sidechains are predicted rather unambiguously. In our datasets, accurate predictions defined as the correct amino acid type being the most probable reached 89% (Fig. 1c).

Next, sequential connectivity information is obtained by matching the ^{13}C–^{1}H frequencies of every TOCSY AAIG (*i*-1) to ^{13}C–^{1}H frequencies present in any other NOESY AAIG (*i*), excluding the NOESY AAIG with the same root resonance (^{15}N–^{1}H) as the TOCSY AAIG (Supplementary Figure 3a). The sequential connectivities established for each TOCSY AAIG may vary in occupancy rate, defined as the ratio of matched frequencies versus the total number of TOCSY frequencies (Supplementary Figure 3a). The algorithm creates a directed rooted

tree from each AAIG and adds progressively edges and nodes using the connectivity information, until it reaches a maximum chain length (Supplementary Figure 3b). As a tradeoff between efficiency and memory consumption, maximum length is set to six. Each chain X is then assigned a probability of occurrence given by the product of the probabilities of each connectivity type in that chain as shown in Eq. 3 to estimate the significance of each chain

$$P(X) = \prod_{k=1}^{L-1} P(X_{k\to k+1}) = \prod_{k=1}^{L-1} P(\text{occupancy rate}) \quad (3)$$

where L is the chain length and k the position in the chain. In principle, chains with higher occupancy rate of connectivities are more likely to be correct.

Subsequently, the chains are used to generate a larger number of peptide sequence segments using the amino acid-type predictions obtained earlier. Each peptide is aligned to the protein sequence using the Needleman–Wunsch algorithm[47]. Many peptides are discarded at this stage due to alignment mismatches. For the aligned peptides, an alignment score S_{align} is computed using the BLOSUM90 similarity matrix[48], which quantifies the importance of the alignment to a specific amino acid sequence. Taken all the above into account, the weighted probability of assigning an AAIG from chain X to a specific residue in the protein sequence is given by Eq. 4.

$$S(X) = P(\text{AA}|\text{CCS}) * P(X) * S_{\text{align}} \quad (4)$$

Since several different chains can be mapped at overlapping positions in the protein sequence, multiple AAIGs may correspond to each protein residue. For each position in the protein sequence, a confidence score (C_s) is computed for every AAIG corresponding to the given position by summing over all the chains as indicated in Eq. 5.

$$C_s = \sum^{\text{all chains } X} S(X) \quad (5)$$

In order to identify the correct AAIGs mapped to the protein sequence from the large pool of aligned chains, the 4D-CHAINS algorithm exploits the overlap information by performing OLC assembly similar to DNA assembly techniques[22] (Fig. 2). From N- to C-terminus of the protein sequence, series of aligned chains are merged to contigs with identical overlap of length L-1, where L is the length of the chains. A contig terminates if there is no overlap to extend or if it encounters an AAIG that is already part of it. Chains that cannot be merged to contigs are considered spurious and discarded. Finally, all contigs are aligned to the protein sequence. For an AAIG to be assigned to a given residue in the protein sequence, two conditions must be met. First, only the absolute consensus AAIGs among the different contigs are taken into account for a given position in the sequence and, second, the consensus AAIG must have the higher confidence score for the given position (Fig. 2).

Mapping of AAIGs to the protein sequence is accomplished by a succession of iterations that differ in three parameters used: (i) the length of chains built; (ii) a Z-score cutoff that controls the amino acid-type predictions to be considered per TOCSY AAIG when chains are translated to peptides; and (iii) the occupancy rate of connectivities between a TOCSY AAIG and all matched NOESY AAIGs.

In the first iteration, stringent criteria are applied to ensure greater fidelity of predictions and extract long chains ($L = 6$) that are less likely to be aligned in a wrong position of the sequence. OLC assembly provides an additional level of scrutiny and removes lonely chains that cannot be extended to either end and are likely false. Accordingly, only consensus AAIGs are selected and if there is agreement with the probabilistic model (confidence score), then are mapped to certain positions of the protein sequence. Mapped AAIGs are restrained in successive rounds by eliminating all amino acid-type predictions and connectivities they participate in that are inconsistent with the NH mapping. This reduces noise and allows us to proceed gradually with shorter chains (minimum length 3) to fill short regions in the sequence that are flanked by gaps in connectivities or proline residues, incorporate AAIGs in the sequence that fulfill the connectivity criteria but have low amino acid-type probability due to abnormal chemical shifts, and finally account for the fact that NOESY AAIGs may not match all frequencies of a TOCSY AAIG. In each round both the OLC and the probabilistic rule must be met for accepting additionally mapped AAIGs to be restrained in the following round.

In the present implementation of the 4D-CHAINS algorithm, no mapping mistakes were made for the four protein targets. The NH-mapping coverage varied between 96 and 100% (Supplementary Figure 4; left column). To better evaluate the mapping performance of 4D-CHAINS, only the ^{13}C-^{1}H correlated frequencies of α- and β-atoms were retained in the TOCSY input peak list to imitate the scenario of using a 4D CBHBCAHA(CO)NH experiment in conjunction with the 4D ^{13}C,^{15}N edited HMQC-NOESY-HSQC. Interestingly, only the coverage dropped slightly but again no mistakes were introduced (Supplementary Figure 4; right column). This control experiment highlights the robustness of 4D-CHAINS that stems mainly from the predictive power of ^{13}C–^{1}H correlated chemical shifts and the reliability of the connectivities established when carbon and proton frequencies are coupled.

For all AAIGs mapped to the protein sequence 4D-CHAINS obtains the assignment of aliphatic atoms by matching the observed ^{13}C–^{1}H correlated frequencies to their distributions in the 2D probability density maps. First, the TOCSY frequencies are assigned to atom types of the previous amino acid in the sequence. Based on the amino acid type, pairs of ^{13}C–^{1}H frequencies that differ by 0.2 p.p.m. or less in the carbon frequency are grouped to methylene moieties. The atom-type probability for these moieties is taken as the logarithmic average of the individual probabilities. Accordingly, all combinations of permutations are computed and the permutation with the highest probability provides the atom-type assignments for the TOCSY observed frequencies. It has been reported before[49] that automated assignments based on TOCSY-type transfer may interchange between atoms of certain amino acids because their chemical shift distributions overlap partially, as seen in the 2D probability density maps of correlated chemical shifts (Supplementary Figure 2). Another source of erroneous assignments may result from incomplete TOCSY patterns that become common as the size of the protein increases and also depend on the isotropic mixing period. For instance, in Leu residues, often the TOCSY observed correlations correspond to the α, β and one of the isopropyl atoms. Due to the missing correlations, any of the methyl groups could be wrongly assigned to the γ atom and vice versa, depending on the observed chemical shifts. 4D-CHAINS assigned all TOCSY correlations with an average error rate of 0.2% for 1845 ^{13}C–^{1}H moieties in total (4 errors out of 1845 types) (Supplementary Figures 6 and 7). Minor TOCSY-based misassignments should have in principle little impact on structure calculations driven by long-range NOEs because they involve intraresidue atoms.

Next, all TOCSY-derived assignments are transferred to the NOESY spectra starting from the last residue and going backwards. For every amino acid, first the intraresidue NOE peaks are assigned (i) by matching the TOCSY assigned peaks of the successive residue, and then any sequential NOE peaks (i-1) by matching its own TOCSY assigned peaks. It has been noted from the early days of NMR[1], and supported later by inter-proton statistics[21], that for any given amide the observed NOE correlations to aliphatic protons are predominantly intraresidue (i) and sequential (i-1). As a proof of that, 4D-CHAINS traced 99% of TOCSY peaks as intraresidue NOE correlations and 89% as sequential NOE correlations. This analysis demonstrates that for any given amide and its successive one, a large portion of common NOEs they share correspond to the aliphatic atoms of the former. This is particularly true for the methyl groups that in principle yield strong NOE correlations to amides, both intraresidue and sequential. The only exception is the distant methyl group of Met. Intraresidue NOE correlations are uniformly present to any type of secondary structure (α-helix, β-sheet, loops), whereas sequential NOE correlations are most prevalent in β-sheets. Yet, most of the missing sequential NOE correlations correspond to certain atom types (δ and ε of Lys, δ of Arg, ε of Met, and to a lower extent γ of Leu and γ1 of Ile).

Common NOEs are utilized to derive missing TOCSY-based assignments. For every residue separately, the NOESY peak intensities are normalized and peaks with low intensity (threshold 0.1 or specified otherwise) are left out. 4D-CHAINS scans the sequence backwards. For each residue where there is a missing assignment, it matches its unassigned NOE peaks with the unassigned NOE peaks of the next residue. For each peak where there is a match, a probability is derived for the missing atom-type assignments from the 2D density map of the particular residue type. Atom-type probabilities must belong to the 80th percentile of the density maps to be considered further. This filter prevents making any decisions when the correct assignment does not belong to any of the matched peaks. Accordingly, each probability is modified by the intensity of the corresponding peak to a score. This process is necessary to identify the correct assignment of methyl groups among the available options, because intraresidue methyl-amide NOEs yield stronger correlations. Several intensity transformations were tested extensively (see below) and the best performance for obtaining NOESY-type methyl assignments is given by the product of the 2D histogram probability and the intensity of the NOESY peak transformed by an exponential function, e.g., 2Dprob×(100×intensity^2). The highest score or product of scores provides the assignments for the missing atom types.

The efficiency of 4D-CHAINS in obtaining atom-type assignments from the 4D-HCNH NOESY spectrum has been tested in three different scenarios (Supplementary Figure 5 and Supplementary Table 1). In the current workflow 4D-CHAINS sought assignments not present in the TOCSY spectra. For the four different datasets, it assigned 13% of additionally assignable aliphatic atoms (277 carbon types) with an average error rate of 8.7% (24 errors out of 277 types). Then, it operated on synthetic data of a 4D CBHBCAHA(CO)NH experiment. It performed NH mapping successfully, assigned correctly all α- and β-atoms (1253 carbon types or 56.2% of all assignable atoms) and completed the missing assignments from the NOESY spectrum, where it assigned 37.9% of additionally assignable atoms (847 carbon types) with an average error rate of 5.1% (43 errors out of 847 types). Finally, only the backbone amide ^{15}N,^{1}H HSQC frequencies were provided and 4D-CHAINS was asked to assign all aliphatic atoms from the 4D-HCNH NOESY spectrum (Supplementary Figure 8). 4D-CHAINS was able to assign 91.1% of all assignable aliphatic atoms (2033 carbon types) with an average error rate of 5.5% (112 errors out of 2033 types). In all cases the assignment error rate for methyl groups was lower: 7.9% for the first scenario (6 errors out of 76 methyls assigned), 3.8% for the second scenario (15 errors out of 393 methyls assigned) and 3.3% for the third scenario (15 errors out of 457 methyls assigned).

Automated atom assignments using FLYA. For FLYA resonance assignment calculations all available spectra were used as input, that is, 4D-HCNH TOCSY, 4D-HCNH NOESY and 4D-HCCH NOESY. The calculations were performed using the demo script provided with CYANA distribution shown in the Supplementary Methods.

NOE assignment and structure determination using Rosetta. NOE-based structural ensembles were generated using the csrosetta3 toolbox integrated within the Rosetta3 software suite. autoNOE-Rosetta[19] is one of the protocols included in the toolbox which performs automatic assignment of NOEs and structure determination based on the highly parallel RASREC-Rosetta[50] conformational sampling engine, which can successfully determine well-converged structures from sparse NMR data[19, 51]. The main principle is to iterate the NOE assignment algorithm alongside a multi-stage (I–VIII) conformational sampling process, towards obtaining a network of long-range NOEs that drive structure refinement to the global minimum of the Rosetta energy function[24]. The protocol uses as input initial assignments of NOE cross-peaks, derived from the chemical shift lists provided by 4D-CHAINS. The selection of high-ranking initial assignments of NOE cross-peaks depends on several factors, including: symmetry of the peaks, chemical shift matching score and network anchoring. Long-range NOE restraints and backbone chemical shift fragments guide the generation of batches of preliminary, low-resolution structures, that are in turn utilized to evaluate and refine the NOE assignments. Short and medium-range NOEs are also assigned by the program, but not utilized in structure refinement. The autoNOE-Rosetta further eliminates peaks during the sampling process. In the final stages, only highly converged, lowest-energy models that satisfy the maximum number of assigned NOE distance restraints are retained.

In practice, the process of setting up the protocol and generating models involves the following steps. (1) Preparation of chemical shift, NOE peaks and sequence files. From the 4D-CHAINS XEASY chemical shift table, we generate a TALOS[52] file and perform empirical prediction of backbone torsion angles using TALOS-N[53] for a given protein sequence. Based on the predicted chemical shift order parameter, we retain only the rigid regions of the structure. (2) Fragment selection[54] from high-resolution structures in the PDB[23]. We use the TALOS-N φ, ψ and secondary structure-type predictions to bias the selection of 3- and 9-residue backbone fragments, excluding fragments derived from homologs to the target sequence present in the database. (3) Automated setup of autoNOE-Rosetta calculations for a range of restraint weight values.

According to this general procedure, we performed two sets of calculations for each target using NOE peak lists that included either (i) amide to aliphatic (HCNH) only, or (ii) amide to aliphatic and aliphatic to aliphatic (HCNH +HCCH). To improve sampling for nEIt, NOEs were supplemented by one RDC dataset. All calculations were setup with standard restraint weights of 5, 10, 25 and 50. The optimum restraint weight was selected based on an empirical cost function that considers individual restraint weights, Rosetta all-atom energies (talaris2014.wts[24]), and degree of structural convergence in each calculation. Finally, we select an ensemble of 10 lowest-energy structures that show minimum NOE violations. A detailed method to setup the calculations and analyze the models is available in the Supplementary Methods.

NOE assignment and structure determination using CYANA. The 3D structural ensembles of all four target proteins used in this study were calculated using the CYANA[27] software suite, supplied with the same input datasets as with autoNOE-Rosetta. Depending on the size of each protein target, CYANA calculations required 45–90 min on 4 CPUs. The script with all parameters for CYANA calculations is available in the Supplementary Methods.

Ensemblator analysis. Analysis of the ensembles was performed using the Ensemblator[25, 26] software for atom- and residue-level global and local comparisons. The Ensemblator first iteratively overlays pairs of structures and finally defines a "common core" of atoms that are consistently within a specified cutoff distance. For each set of comparisons, the needed cutoff distance was automatically determined by the Ensemblator to yield 20–40% of the atoms in the common core. These comparisons also yield pairwise weighted distance metrics that are used to embed the models into an N-dimensional space where N is the number of models. Also, for any specified group of models, an exemplar was defined as the model having the shortest average distance to all other models in its own group. Global comparisons between groups are performed after the common core atoms are used to overlay structures and local backbone comparisons are calculated based on the locally overlaid dipeptide residual which converts φ, ψ differences to a single distance[25]. The global and local comparisons involve quantifying the levels of variation for each residue within and between defined groups so that the level of intragroup variation can be compared with the intergroup variation.

Finally, the models in the crystallographic ensemble consisting of 51 aLP structures, used for Ensemblator analysis, were obtained from the CoDNaS[28] database by searching for α-lytic protease and utilizing all the available X-ray structures.

Restraint violation analysis. NOE restraint violations among the 10 lowest-energy models in each calculation are reported separately for different classes of restraints assigned by autoNOE-Rosetta (Supplementary Tables 2, 6–9). First, restraints are automatically divided into three confidence classes according to a total assignment probability score[19]: highly confident (HI) (probability >70%), confident (MED) (probability >45%) and least confident (LOW) (probability <45%). Second, the ambiguity score reflecting assignment uniqueness[19] further classifies constraints into ambiguous (AMBIG) (ambiguity score >0.1), near unambiguous (NEAR_UNAMBIG) (ambiguity score <0.1) or unambiguous (UNAMBIG) (ambiguity score <0.01). Therefore, according to these criteria, each constraint can be classified into one of the six classes: HI_UNAMBIG, HI_NEAR_UNAMBIG, HI_AMBIG, MED_UNAMBIG, MED_AMBIG and LOW_AMBIG. Finally, due to the lack of stereo-specific assignments by 4D-CHAINS, the resulting autoNOE restraints are structurally degenerate and are therefore treated using an effective distance computed as the r^{-6} average between all the possible pairs of atoms[55]. We used a 7 Å upper distance bound to identify violations in the resulting structurally degenerate NOE restraints, shown as an average over the 10 lowest-energy models in each structural ensemble. The choice of 7 Å as upper distance bound is attributed to the use of a 70 ms mixing time where through-space magnetization transfer between closer protons can happen within a maximum distance range of 7 Å[56]. This statement was found to be true by direct observation of distances corresponding to confidently assigned NOEs in the X-ray structure of aLP (PDB ID 1PO1).

Computational cost. 4D-CHAINS takes an average of half an hour to run on a commodity computer. All autoNOE-Rosetta structure calculations were performed at the UCSC Baker cluster with 13 compute nodes (AMD Opteron(tm), 2.4 GHz Processor 6378) and 32 cores per compute node. Typical message passing interface calculations are run in parallel on 100 cores, and depending on target size, take an average of: (i) 6–8 h (150 aa), (ii) 12–14 h (200 aa) and (iii) 16–18 h (250 aa). A total of approximately 2 million CPU hours was used for the various development stages of the method.

CS-Rosetta support for NMR Exchange Format. NMR restraint datasets are now represented using a new open standard, NMR Exchange Format (NEF)[57]. NEF is a self-contained format designed to be machine readable by common NMR structure determination software tools. The file is divided into sections where each section corresponds to the data used for structure calculation. Full specification of each section in NEF can be found at: https://github.com/NMRExchangeFormat/NEF/blob/master/specification/Overview.md.

NEF provides support for a set of identifiers to be used by software tools. We utilize the identifiers provided by the NEF specifications to design NEF converter and NEF parser tools as part of the csrosetta3 toolbox. NEF converter is a tool that can take sequence information, chemical shift assignments, RDC data (if available), distance restraints and peak information used for structure calculation and convert it to the standard NEF file format for deposition in databases (that support NEF).

Similarly, we also provide a series of tools to extract respective information from NEF file into FASTA, NOE restraint, chemical shift and peak files for subsequent automatic setup of CS-Rosetta for structure calculations. See supplementary information for detailed commands to convert to NEF and parse NEF file.

wwPDB data deposition. Currently, wwPDB does not support the NEF format, and we therefore utilized the NEF to BMRB translator program provided by BMRB[58] (https://github.com/kumar-physics/BMRBTranslator) to convert from NEF to NMR-STAR format for data deposition. The deposited NMR-STAR file consists of chemical shifts, peaks and RDCs used for structure calculation.

Code availability. 4D-CHAINS is available on github (https://github.com/tevang/4D-CHAINS) for non-commercial usage. The updated CS-Rosetta (version 3.4) software and the detailed documentation for installation and usage can be obtained at the CS-Rosetta web server (https://csrosetta.chemistry.ucsc.edu). The current version (3.4) of CS-Rosetta also supports conversion of data to NMR Exchange Format for deposition to the wwPDB.

Data availability. Biological Magnetic Resonance Bank: chemical shifts, peak lists, RDCs have been deposited under 30322, 30325, 30326, and 30327 BMRB codes. Protein Data Bank: restraint lists and coordinates have been deposited under 5WOT, 5WOX, 5WOY, and 5WOZ PDB codes. Other data are available from the corresponding authors upon reasonable request.

Received: 20 August 2017 Accepted: 12 December 2017
Published online: 26 January 2018

References

1. Wüthrich, K. *NMR of Proteins and Nucleic Acids*. (Wiley: New York, 1986).
2. Ikura, M., Kay, L. & Bax, A. A novel approach for sequential assignment of proton, carbon-13, and nitrogen-15 spectra of larger proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to Calmodulin. *Biochemistry* **29**, 4659–4667 (1990).

3. Kay, L. E., Ikura, M., Tschudin, R. & Bax, A. Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. *J. Magn. Reson.* **89**, 496–514 (1990).
4. Güntert, P. Automated structure determination from NMR spectra. *Eur. Biophys. J.* **38**, 129–143 (2009).
5. Schmidt, E. & Güntert, P. A new algorithm for reliable and general NMR resonance assignment. *J. Am. Chem. Soc.* **134**, 12817–12829 (2012).
6. Bahrami, A., Assadi, A. H., Markley, J. L. & Eghbalnia, H. R. Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy. *PLoS Comput. Biol.* **5**, e1000307 (2009).
7. Guerry, P., Duong, V. & Herrmann, T. CASD-NMR 2: robust and accurate unsupervised analysis of raw NOESY spectra and protein structure determination with UNIO. *J. Biomol. NMR* **62**, 473–480 (2015).
8. Kay, L., Clore, G., Bax, A. & Gronenborn, A. Four-dimensional heteronuclear triple-resonance NMR spectroscopy of interleukin-1 beta in solution. *Science* **249**, 411–414 (1990).
9. Kazimierczuk, K. & Orekhov, V. Non-uniform sampling: post-Fourier era of NMR data collection and processing. *Magn. Reson. Chem.* **53**, 921–926 (2015).
10. Trautwein, M., Fredriksson, K., Moller, H. M. & Exner, T. Automated assignment of NMR chemical shifts based on a known structure and 4D spectra. *J. Biomol. NMR* **65**, 217–236 (2016).
11. Pritišanac, I. et al. Automatic assignment of methyl-NMR spectra of supramolecular machines using graph theory. *J. Am. Chem. Soc.* **139**, 9523–9533 (2017).
12. Shen, Y. et al. Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl. Acad. Sci. USA* **105**, 4685–4690 (2008).
13. Cavalli, A., Salvatella, X., Dobson, C. M. & Vendruscolo, M. Protein structure determination from NMR chemical shifts. *Proc. Natl. Acad. Sci. USA* **104**, 9615–9620 (2007).
14. Raman, S. et al. NMR structure determination for larger proteins using backbone-only data. *Science* **327**, 1014–1018 (2012).
15. Lange, O. F. et al. Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proc. Natl. Acad. Sci. USA* **109**, 10873–10878 (2012).
16. Thompson, J. M. et al. Accurate protein structure modeling using sparse NMR data and homologous structure information. *Proc. Natl. Acad. Sci. USA* **109**, 9875–9880 (2012).
17. Shen, Y. & Bax, A. Homology modeling of larger proteins guided by chemical shifts. *Nat. Methods* **12**, 747–750 (2015).
18. Tang, Y. et al. Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nat. Methods* **12**, 751–754 (2015).
19. Lange, O. F. Automatic NOESY assignment in CS-RASREC-Rosetta. *J. Biomol. NMR* **59**, 147–159 (2014).
20. Zhang, Z., Porter, J., Tripsianes, K. & Lange, O. F. Robust and highly accurate automatic NOESY assignment and structure determination with Rosetta. *J. Biomol. NMR* **59**, 135–145 (2014).
21. Xu, Y., Zheng, Y., Fan, J. & Yang, D. A new strategy for structure determination of large proteins in solution without deuteration. *Nat. Methods* **3**, 931–937 (2006).
22. Li, Z. et al. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief. Funct. Genom.* **11**, 25–37 (2012).
23. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
24. O'Meara, M. J. et al. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *J. Chem. Theory Comput.* **11**, 609–622 (2015).
25. Clark, S. A., Tronrud, D. E. & Karplus, P. A. Residue-level global and local ensemble-ensemble comparisons of protein domains. *Protein Sci.* **24**, 1528–1542 (2015).
26. Brereton, A. E. & Karplus, P. A. Ensemblator v3: Robust atom-level comparative analyses and classification of protein structure ensembles. *Protein Sci.* **27**, 41–50 (2018).
27. Güntert, P. & Buchner, L. Combined automated NOE assignment and structure calculation with CYANA. *J. Biomol. NMR* **62**, 453–471 (2015).
28. Monzon, A.M., Rohr, C.O., Fornasari, M.S. & Parisi, G. CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. *J. Biol. Databases Curation.* https://doi.org/10.1093/database/baw038 (2016).
29. Jasnovidova, O., Krejcikova, M., Kubicek, K. & Stefl, R. Structural insight into recognition of phosphorylated threonine-4 of RNA polymerase II C-terminal domain by Rtt103p. *EMBO Rep.* **18**, 906–913 (2017).
30. Jasnovidova, O. et al. Structure and dynamics of the RNAPII CTDsome with Rtt103. *Proc. Natl. Acad. Sci. USA* **114**, 11133–11138 (2017).
31. Frueh, D. P. Practical aspects of NMR signal assignment in larger and challenging proteins. *Prog. Nucl. Magn. Reson. Spectrosc.* **78**, 47–75 (2014).
32. Kainosho, M. et al. Optimal isotope labelling for NMR protein structure determinations. *Nature* **440**, 52–57 (2006).
33. Tugarinov, V., Kanelis, V. & Kay, L. E. Isotope labeling strategies for the study of high-molecular-weight proteins by solution NMR spectroscopy. *Nat. Protoc.* **1**, 749–754 (2006).
34. Vinarov, D. A. & Markley, J. L. High-throughput automated platform for nuclear magnetic resonance–based structural proteomics. *Expert. Rev. Proteom.* **2**, 49–55 (2005).
35. Dias, D. M. & Ciulli, A. NMR approaches in structure-based lead discovery: recent developments and new frontiers for targeting multi-protein complexes. *Prog. Biophys. Mol. Biol.* **116**, 101–112 (2014).
36. Kazimierczuk, K., Zawadzka, A., Koźmiński, W. & Zhukov, I. Random sampling of evolution time space and Fourier transform processing. *J. Biomol. NMR* **36**, 157–168 (2006).
37. Stanek, J., Augustyniak, R. & Koźmiński, W. Suppression of sampling artefacts in high-resolution four-dimensional NMR spectra using signal separation algorithm. *J. Magn. Reson.* **214**, 91–102 (2012).
38. Delaglio, F. et al. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. Nmr.* **6**, 277–293 (1995).
39. Goddard, T. D. & Kneller, D. G. *SPARKY 3* (University of California, San Francisco, San Francisco, 2000).
40. Bax, A., Kontaxis, G. & Tjandra, N. Dipolar couplings in macromolecular structure determination. *Methods Enzymol.* **339**, 127–174 (2001).
41. Hansen, M. R., Mueller, L. & Pardi, A. Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. *Nat. Struct. Mol. Biol.* **5**, 1065–1074 (1998).
42. Fitzkee, N. C. & Bax, A. Facile measurement of 1H–15N residual dipolar couplings in larger perdeuterated proteins. *J. Biomol. NMR* **48**, 65–70 (2010).
43. Gronwald, W. & Kalbitzer, H. R. Automated structure determination of proteins by NMR spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.* **44**, 33–96 (2004).
44. Baran, M. C., Huang, Y. J., Moseley, H. N. B. & Montelione, G. T. Automated analysis of protein NMR assignments and structures. *Chem. Rev.* **104**, 3541–3556 (2004).
45. Reiping, W. & Vranken, W. F. Validation of archived chemical shifts through atomic coordinates. *Proteins* **78**, 2482–2489 (2010).
46. Marin, A., Malliavin, T. E., Nicolas, P. & Delsuc, M. A. From NMR chemical shifts to amino acid types: Investigation of the predictive power carried by nuclei. *J. Biomol. NMR* **30**, 47–60 (2004).
47. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
48. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919 (1992).
49. Hiller, S., Joss, R. & Wider, G. Automated NMR assignment of protein side chain resonances using automated projection spectroscopy (APSY). *J. Am. Chem. Soc.* **130**, 12073–12079 (2008).
50. Lange, O. & Baker, D. Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. *Proteins* **80**, 884–895 (2012).
51. Warner, L. R. et al. Structure of the BamC two-domain protein obtained by Rosetta with a limited NMR data set. *J. Mol. Biol.* **411**, 83–95 (2011).
52. Cornilescu, G., Delaglio, F. & Bax, A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13**, 289–302 (1999).
53. Shen, Y. & Bax, A. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J. Biomol. NMR* **56**, 227–241 (2013).
54. Gront, D., Kulp, D. W., Vernon, R. M., Strauss, C. E. M. & Baker, D. Generalized fragment picking in Rosetta: design, protocols and applications. *PLoS ONE* e23294 (2011).
55. Nilges, M. Ambiguous distance data in the calculation of NMR structures. *Fold Des.* **2**, S53–S57 (1997).
56. Foster, M. P., McElroy, C. A. & Amero, C. D. Solution NMR of large molecules and assemblies. *Biochem. (Mosc.).* **46**, 331–340 (2007).
57. Gutmanas, A. et al. NMR Exchange Format: a unified and open standard for representation of NMR restraint data. *Nat. Struct. Mol. Biol.* **22**, 433–434 (2015).
58. Ulrich, E. L. et al. BioMagResBank. *Nucleic Acids Res.* **36**, D402–D408 (2008).

Acknowledgements

We thank Peter Lukavsky, Richard Stefl and Arie Geerlof for providing NMR samples and Alison Barrett for assistance during the early stages of developing the csrosetta3 code. This research was supported by a grant from the Czech Science Foundation (15-22380Y), project CEITEC 2020 (LQ1601) with financial contribution from the MEYS CR and National Programme for Sustainability II, and a Marie Curie Career Integration Grant (618223) to K.T., by NIH grant R01GM083136 to P.A.K., a K-22 Career Development and an R35 Outstanding Investigator Award to N.G.S. through NIAID(AI2573-

01) and NIGMS(1R35GM125034-01), respectively. CIISB research infrastructure project LM2015043 funded by MEYS CR is gratefully acknowledged for the financial support of the measurements at CEITEC Josef Dadok National NMR Centre. We acknowledge the UCSC 800 MHz NMR facility supported by the Office of the Director, NIH, under High End Instrumentation (HIE) Grant S10OD018455.

Author contributions

K.T. conceived and together with N.G.S. designed the project. T.E. and K.T. developed and executed the 4D-CHAINS algorithm. J.N. and K.T. developed, recorded and analyzed 4D NMR experiments. S.N. and N.G.S. developed the CS-Rosetta3 software and performed parallel structure calculations, structure analysis and validation, additional testing of 4D-CHAINS and NMR experiments. A.E.B. and P.A.K. performed Ensemblator analysis. R.R.D. and V.V. prepared labeled protein samples and recorded RDC data for nEIt. N.G.S., S.N., T.E. and K.T. wrote the paper, with feedback from all authors.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-017-02592-z.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018

ARTICLES
https://doi.org/10.1038/s41594-018-0141-6

nature structural & molecular biology

De novo design of a non-local β-sheet protein with high stability and accuracy

Enrique Marcos[1,2,3,10*], Tamuka M. Chidyausiku[1,2,10], Andrew C. McShan[4], Thomas Evangelidis[5], Santrupti Nerli[4,6], Lauren Carter[1,2], Lucas G. Nivón[1,2,7], Audrey Davis[1,2,8], Gustav Oberdorfer[1,2,9], Konstantinos Tripsianes[5], Nikolaos G. Sgourakis[4] and David Baker[1,2*]

β-sheet proteins carry out critical functions in biology, and hence are attractive scaffolds for computational protein design. Despite this potential, de novo design of all-β-sheet proteins from first principles lags far behind the design of all-α or mixed-αβ domains owing to their non-local nature and the tendency of exposed β-strand edges to aggregate. Through study of loops connecting unpaired β-strands (β-arches), we have identified a series of structural relationships between loop geometry, side chain directionality and β-strand length that arise from hydrogen bonding and packing constraints on regular β-sheet structures. We use these rules to de novo design jellyroll structures with double-stranded β-helices formed by eight antiparallel β-strands. The nuclear magnetic resonance structure of a hyperthermostable design closely matched the computational model, demonstrating accurate control over the β-sheet structure and loop geometry. Our results open the door to the design of a broad range of non-local β-sheet protein structures.

β-sheet protein domains are ubiquitous in nature, carrying out a wide range of functions: transporting hydrophobic molecules, recognition and enzymatic processing of carbohydrates, and scaffolding of virus capsids and antibodies, among others. Although β-sheet protein scaffolds are well suited for incorporating new functions, their design from first principles remains an outstanding challenge. Recent progress in de novo protein design has enabled the accurate design of many hyperstable and structurally diverse proteins, but so far, other than short β-sheet peptides[1–3], all exhibit either all-α or mixed-αβ folds[4]. The design of these last has been considerably facilitated by the derivation of a set of rules describing constraints on the backbone geometry of the loops connecting secondary structure elements[5], but all-β proteins contain additional features that are less well understood. All-β-sheet structures are particularly challenging to design from scratch[6] because a larger fraction of the interactions are non-local (that is, between residues distant along the linear sequence), leading to slower folding rates[7], and because β-strands, particularly at the edges of β-sheets, can aggregate into amyloid-like structures. Hence, few β-sheet protein design studies have sought to generate new backbone structures[8,9] and, except for a recent β-barrel structure with primarily local strand pairings[10], those designs confirmed by high-resolution structure determination have relied heavily on sequence information[11,12] and backbone structures[13,14] from naturally occurring β-sheet proteins.

So far, the de novo design of β-sheet loop connections has been limited to β-hairpins (two antiparallel β-strands interacting via backbone hydrogen bonding and connected through a loop), which is the most local strand pairing possible and, in principle, the fastest to fold. However, these structures lack a critical feature of non-local globular all-β structures: loops connecting β-strands not paired to each other, also known as β-arches[15]. These loops connect distinct β-sheets and pair β-strands with larger sequence separation, and they are essential for enabling the protein fold complexity observed in antibodies, β-solenoids, jellyrolls and Greek key-containing structures generally. Here we set out to identify the general principles for designing non-local β-sheet structures.

Results

Constraints on β-arch geometry. We undertook investigation of the constraints on the backbone geometry of β-strands and connecting loops that arise from hydrogen bonding and the requirement for a compact hydrophobic core. We studied side chain directionality patterns of the two β-strand residues adjacent to β-arch loops (Fig. 1a, left) in naturally occurring protein structures, defining the side chain orientation of the β-strand residue preceding the loop as 'concave' (represented by ↓) if its CαCβ vector is parallel to the vector *d* from the first to the second β-strand, and 'convex' (represented by ↑) if the CαCβ vector is antiparallel to *d*. For the residue following the loop, the side chain pattern is described in the same way, but instead using the vector from the second to the first β-strand (−*d*) as a reference (Fig. 1a). This results in four possible β-arch loop side chain orientation patterns: ↑↑, ↑↓, ↓↑ and ↓↓. We analyzed the side chain patterns and the local backbone geometry (as described with ABEGO torsion bins[16]) of 5,061 β-arch loops from a non-redundant database of natural protein structures (torsion bins A and B are the α-helix and extended regions; G and E regions are the positive φ angle equivalents of A and B; and O is the *cis*-peptide bond conformation; Supplementary Fig. 1). We found that all four side chain orientation patterns frequently occur, and, in contrast to other types of loop connections (that is, αβ, βα and

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA, USA. [3]Institute for Research in Biomedicine (IRB Barcelona), Barcelona Institute of Science and Technology, Barcelona, Spain. [4]Department of Chemistry and Biochemistry, University of California, Santa Cruz, Santa Cruz, CA, USA. [5]CEITEC—Central European Institute of Technology, Masaryk University, Brno, Czech Republic. [6]Department of Computer Science, University of California, Santa Cruz, Santa Cruz, CA, USA. [7]Present address: Cyrus Biotechnology, Seattle, WA, USA. [8]Present address: Amazon, Seattle, WA, USA. [9]Present address: Institute of Biochemistry, Graz University of Technology, Graz, Austria. [10]These authors contributed equally: Enrique Marcos, Tamuka M. Chidyausiku. *e-mail: emarcos82@gmail.com; dabaker@uw.edu

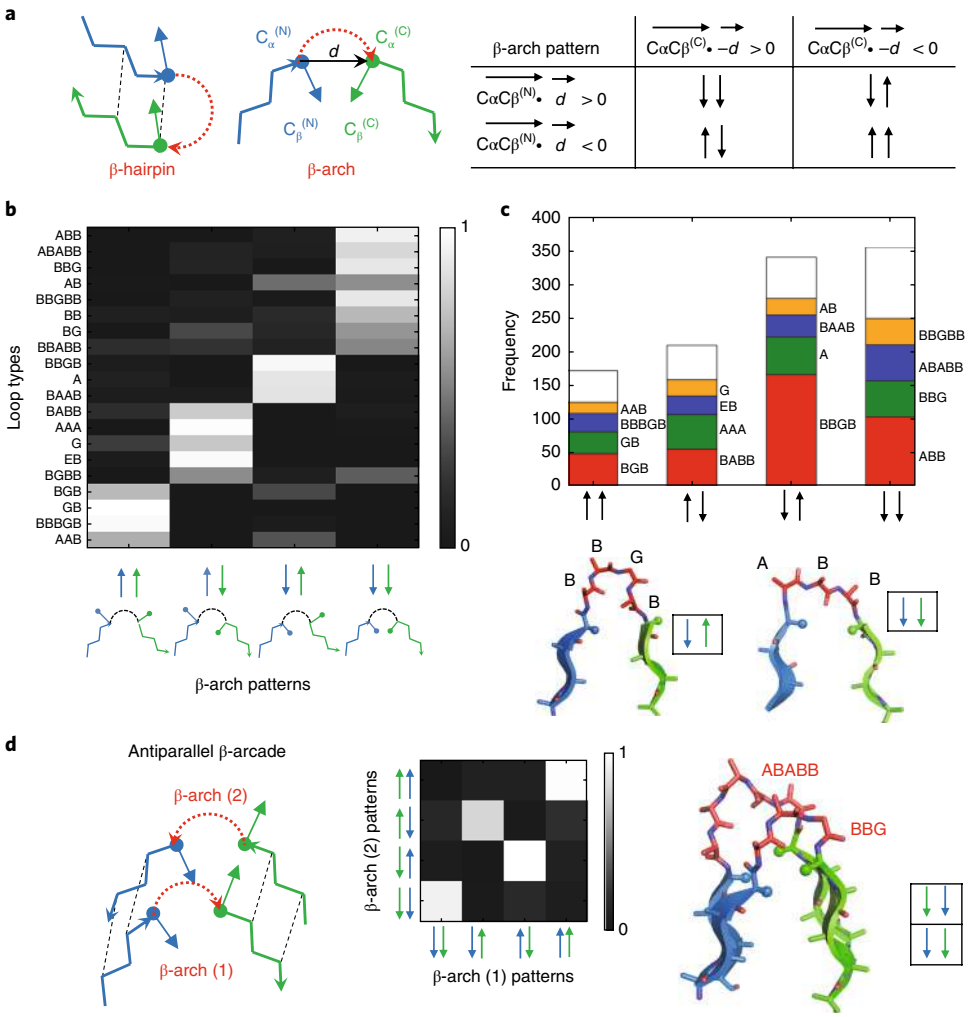


Fig. 1 | Constraints on β-arch geometry. a, Side chain directionality in the β-arch. Comparison between β-hairpin and β-arch (left); the CαCβ and *d* vectors used to define the orientation of the two adjacent side chains are indicated. The four possible side chain directionality patterns are on the right. **b**, Turn type dependence of β-arch side chain patterns. Loops on the *y* axis are described by their ABEGO torsion bins (Supplementary Fig. 1). Most of the loops adopt only one of the four possible side chain patterns. **c**, Frequency of the most common loops for each of the four β-arch side chain patterns. There are strong preferences: for example, BBGB is strongly associated with the ↓↑ pattern, whereas ABB is strongly associated with the ↓↓ pattern (bottom). Only loops with bending <120° (see Methods) and containing between 1 and 5 amino acids were considered in this analysis. **d**, β-arcades consist of two stacked β-arches with in-register strand pairing (left). Because strand pairs of the β-arcade are in register, the side chains adjacent to one β-arch loop must have the same orientation as the paired side chains that are adjacent to the second β-arch loop, and therefore not all loop pairs are allowed (middle). Example of a β-arcade formed by two common β-arches with compatible side chain patterns (right).

β-hairpins)[5], there was no correlation between β-arch loop length and side chain pattern. Instead, each loop ABEGO type, because of the way in which it twists and bends the polypeptide chain[16], is associated with a specific flanking residue side chain pattern (Fig. 1b). The most frequently observed turn types (between 1 and 5 amino acids) for each side chain pattern are listed in Fig. 1c; for example, ABB, BBGB, BABB and BGB are the most frequent loop types for the patterns ↓↓, ↓↑, ↑↓ and ↑↑, respectively.

The next level of non-local interaction complexity in all-β folds involves strand pairing (parallel or antiparallel) between two β-arches forming a β-arcade (Fig. 1d), a common structural motif in naturally occurring β-solenoids[15,17]. Because the β-arch loops are stacked in register, the side chains adjacent to one β-arch loop are likely to have the same orientation as the side chains adjacent to the second β-arch loop; analysis of naturally occurring β-arcades confirms that the side chain patterns of the two β-arch loops are indeed correlated (Fig. 1d, middle).

Jellyroll design principles. The double-stranded β-helix can be regarded as a long β-hairpin wrapped around an axis perpendicular to the direction of β-strands, with β-helical turns formed by the pairing between β-arcades (Fig. 2a). In the compact folded structure, two antiparallel β-sheets pack against each other in a sandwich-like arrangement, with the first strand paired to the last, and all β-strands are connected through β-arch loops except for the central β-hairpin. We aimed to design β-helices with three β-arcades forming two antiparallel four-stranded β-sheets, with the eight β-strands connected through six β-arches and one β-hairpin. The non-local character of the structure grows from the first β-arcade, which starts from the central β-hairpin, to the last one, where the N and C termini are paired.

The analysis from Fig. 1 leads to strong constraints on the construction of β-sheet backbone structures, as the side chain directionality patterns of the β-strands and loops are coupled in several ways. First, the directionality patterns of the loops preceding and following

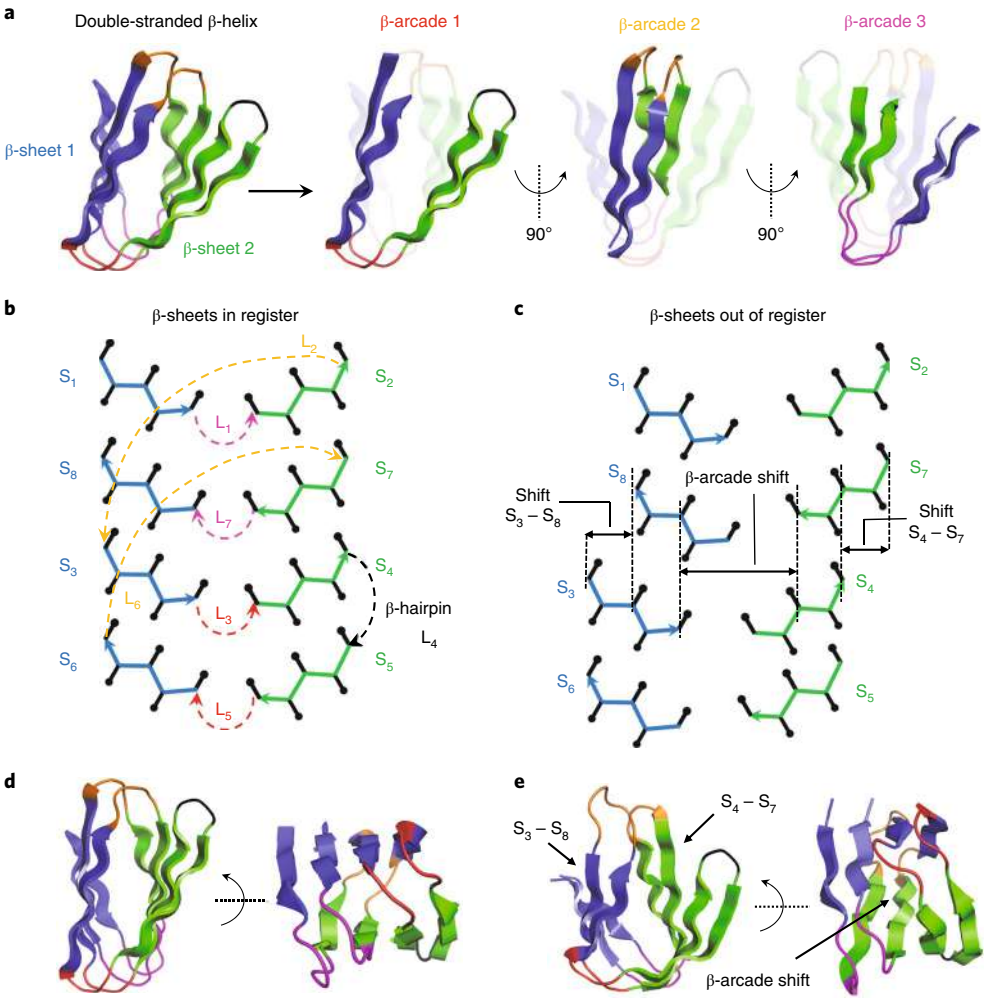


Fig. 2 | Double-stranded β-helix topology specification. a, The double-stranded β-helix fold consists of two four-stranded antiparallel β-sheets (in blue and green) with six β-arch and one β-hairpin connection. Pairs of β-arches forming the three β-arcades are highlighted (right). β-arch loops belonging to the same β-arcade are displayed with the same color throughout the figure (β-arcades 1, 2 and 3 in red, orange and magenta, respectively). **b**, Topology diagram of a designed double-stranded β-helix with all β-strand pairs in register. The Cα traces of the first and second β-sheets are colored in blue and green, respectively. Side chain Cβ positions oriented toward the inner and outer faces of the β-helix are represented with up and down black arrows with rounded tips, respectively. β-arch loops are colored as in **a**. **c**, Definition of β-arcade register shift varied during conformational sampling. The β-arcade register shift (between β-arcades 1 and 3) is determined by the register of β-strand pairs S_3/S_8 and S_4/S_7, and the lengths of β-strands S_3, S_4, S_8 and S_7 (see Methods). In this example, β-strand pairs S_3/S_8 and S_4/S_7 each have a two-residue register shift, resulting in an overall β-arcade register shift of four residues. Loops are omitted to facilitate visualization. **d**, Example of a design model with all β-strand pairs in register forming a sandwich-like structure. **e**, Example of a design model with register shifts between β-arcades 1 and 3 (magenta and red) forming a barrel-like structure.

each β-strand are coupled to the length of the strand (Fig. 2b): for example, a β-strand with an even number of residues that is preceded by a ↑↑ loop must be followed by a ↓↑ or a ↓↓ loop, but not a ↑↑ or ↑↓ loop, owing to the alternating pleating of β-strands. Second, because the β-arcades of the β-helix have paired β-strands and β-arch loops, the side chains adjacent to one β-arch loop must have the same orientation as the paired side chains adjacent to the second β-arch loop (Fig. 1d). Owing to the antiparallel orientation of the β-arcades, ↓↓ and ↑↑ loops are compatible with loops of the same type, but ↑↓ loops are only compatible with ↓↑ loops (Fig. 1d). Third, the twist and curvature of the two β-sheets of the β-helix are constrained by the hydrogen-bonding register between β-arcades 1 and 3 (herein called β-arcade register), and within β-strand pairs S_3/S_8 and S_4/S_7, as shown in Fig. 2c.

De novo design of protein structures. We constructed double-stranded β-helix protein backbones by Monte Carlo fragment assembly using blueprints (representations of the target protein topologies specifying the ordering, lengths and backbone torsion bins of secondary structure elements and loop connections[5]) in conjunction with backbone hydrogen-bonding constraints specifying all pairings between β-strands. We explored strand lengths between 5 and 7 residues and the most commonly observed β-arch loops between 3 and 5 residues (Fig. 1c). The central β-hairpin was designed with two-residue loops following the ββ rule[5]. The register shifts between pairs of β-strands from different β-arcades (1 and 3) were allowed to range from 0 to 2 and the β-arcade register shifts between 0 and 4; strand pairs within the same β-arcade were kept in register. A total of 3,673 combinations were enumerated, of which 1,853 had mutually compatible strand lengths and loop types consistent with the constraints summarized in the previous paragraph. For each of these internally consistent blueprints, we used Rosetta to build thousands of protein backbones. The resulting ensemble of backbone structures has considerable structural diversity; those

with all strands in register had narrow, sandwich-like structures (Fig. 2d), while those with large register shifts had wider, barrel-like structures (Fig. 2e).

For each generated backbone, we carried out flexible-sequence design calculations[18,19] to identify low-energy amino acid identities and side chain conformations providing close complementary packing, side chain–backbone hydrogen bonding in β-arch loops (to pre-organize their conformation and facilitate folding), and high sequence–structure compatibility. We favored inward-pointing charged or polar amino acids at the four edge strands to minimize aggregation propensity[20]. Loop sequences were designed with consensus profiles obtained from fragments with the same backbone ABEGO torsion bins[21]. Because the very large size of the space sampled by our design procedure limits convergence on optimal sequence–structure pairs, we carried out a second round of calculations starting from the blueprints yielding the lowest-energy designs, intensifying sampling at both the backbone and sequence level. For a subset of designs, we introduced disulfide bonds between paired β-strand positions with high sequence separation (for example, between the first and last β-strands) and optimal orientation (see Methods): disulfide bonds distant in primary sequence decrease the entropy of the unfolded state and therefore enhance the thermodynamic stability of the native state. To assess compatibility of the top-ranked designed sequences with their structures, we characterized their folding energy landscape with biased forward folding simulations[21], and those with substantial near-native sampling were subsequently assessed by Rosetta ab initio structure prediction calculations[22,23]. Designs with funnel-shaped energy landscapes (where the designed structure is at the global energy minimum and has a substantial energy gap with respect to alternative conformations) were selected for experimental characterization. Ab initio structure prediction of natural β-sheet proteins tends to oversample local contacts[24,25] (that is, favors β-hairpins over β-arches), but we succeeded in designing sequences with the β-arches sufficiently strongly encoded that they folded in silico to near the designed target structure.

Experimental characterization. For experimental characterization, we chose 19 designs with funnel-shaped energy landscapes ranging between 70 and 94 amino acids (Supplementary Table 1). BLAST searches[26,27] indicated that the designed sequences had little or no similarity with native proteins (lowest Expect (*E*) values ranging from 0.003 to >10; Supplementary Table 2). Synthetic genes encoding the designs (design names are BH_n, where BH stands for β-helix, n stands for the design number and the _ss suffix is used if disulfide bonds are present) were obtained; the proteins were expressed in *Escherichia coli* and purified by affinity chromatography. Sixteen of the designs expressed well and were soluble, and two (BH_10 and BH_11) were monomeric (Supplementary Fig. 2) by size-exclusion chromatography coupled with multi-angle light scattering (SEC-MALS) (most of the non-monomeric designs were either dimers or soluble aggregates). Both monomeric designs had far-ultraviolet circular dichroism spectra (CD) at 25 °C characteristic of β proteins, a melting temperature (T_m) above 95 °C, and well-ordered structures according to 2D ^{1}H-^{15}N heteronuclear single-quantum coherence (HSQC) spectra (Fig. 3a–c and Supplementary Fig. 3). For both designs, the number of NMR peaks matched the number of expected amide resonances based on the protein sequence, but the higher stability of BH_10 under the conditions of the NMR experiments made it a better candidate for NMR structure determination.

The two monomeric designs with well-ordered structures were among those with better-packed cores and a larger proportion of β-arch loops containing prolines and with backbone polar atoms making hydrogen bonds (Supplementary Table 3). β-arch loops that are structurally preorganized with the polar groups making internal hydrogen bonding likely favor folding to the correct topology and contribute to stability by compensating for the loss of interactions with water of polar groups in the side chains and backbone. These interactions likely disfavor the competing local strand-pairing arrangement in which the two strands form a β-hairpin; this is a very common pathology in ab initio structure prediction[25]. For the most stable dimeric design (BH_6), we introduced disulfide bonds to stabilize protein regions having contacts with large sequence separation (for example, between the N- and C-terminal strands), but this did not succeed in yielding stable monomers. Addition of an α-helix to the C terminus (one of the two extremes of the β-helix), as a capping domain protecting the strand edges from intermolecular pairing, also failed to yield stable monomers, even in combination with disulfide bonds. This suggests that the sequence of the core β-sheet must strongly encode its structure independently of disulfide bonds or protecting domains aimed at increasing stability.

NMR structure of a de novo–designed β-helix. We succeeded in solving the structure of BH_10 by 4D NMR spectroscopy (Fig. 3d, Table 1 and Supplementary Fig. 4), using the 4D-CHAINS/AutoNOE-Rosetta automated pipeline for resonance assignments and structure calculation[28], and found it to be in very close agreement with the computational model (Cα r.m.s. deviation (r.m.s.d.) of 0.84 Å, averaged over 10 NMR models). The overall topology is accurately recapitulated, including all strand pairings, register shifts and loop connections, as supported by 132 long-range nuclear Overhauser effects (NOEs) between backbone amide and side chain protons (Supplementary Fig. 5). The designed aliphatic and aromatic side chain packing in the protein core, as well as salt bridge interactions across the two β-sheet surfaces, was also accurately reproduced; three salt bridges between the two paired β-arcades and one within the third β-arcade are well supported by the observed NOEs (Supplementary Fig. 6). The agreement in both the backbone conformation and hydrogen-bonding interactions of the loops forming the three β-arcades is remarkable given that these elements are the most flexible parts of the structure and therefore difficult to design owing to sampling bottlenecks. The β-arcades were designed with pairs of β-arch loops that interact via backbone-backbone hydrogen bonds (owing to the complementarity between their backbone conformations) stabilizing loop pairing and avoiding burial of polar backbone atoms (see Supplementary Fig. 7 for the BH_10 loop sequences and side chain patterns). For example, β-arcade 1 is formed by BBG and ABB loops, and the buried backbone NH group of the G position in the former makes a hydrogen bond with the buried backbone C=O of the neighboring loop (Fig. 3e). The other two β-arcades were designed with one β-arch loop containing buried and fully hydrogen-bonded asparagines (four hydrogen bonds in total) that stabilize both loop pairing and the local β-arch conformation (of ABABB loops). By design, the asparagine side chain geometry was further stabilized with hydrophobic stacking interactions from the two β-arch loops of the same arcade. The high degree of convergence of the designed rotamer in the NMR ensemble illustrates the high structural preorganization of this particular motif (Fig. 3f).

The amino acid sequence of BH_10 is unrelated to any sequence in the NCBI nr database (BLAST found one hit with insignificant sequence similarity; *E* value 6.3). We searched the Protein Data Bank (PDB) for similarities in structure (using the Dali server[29] with the lowest-energy NMR model as the query structure) or sequence (with HHpred[30] for sensitive profile-based sequence search), and identified matches similar in fold but containing additional and irregular secondary structures and longer loops. These matches are all homodimers with sheet-to-sheet interface packing (Supplementary Fig. 8) or domains integrated in larger structures, in sharp contrast to the BH_10 monomer.

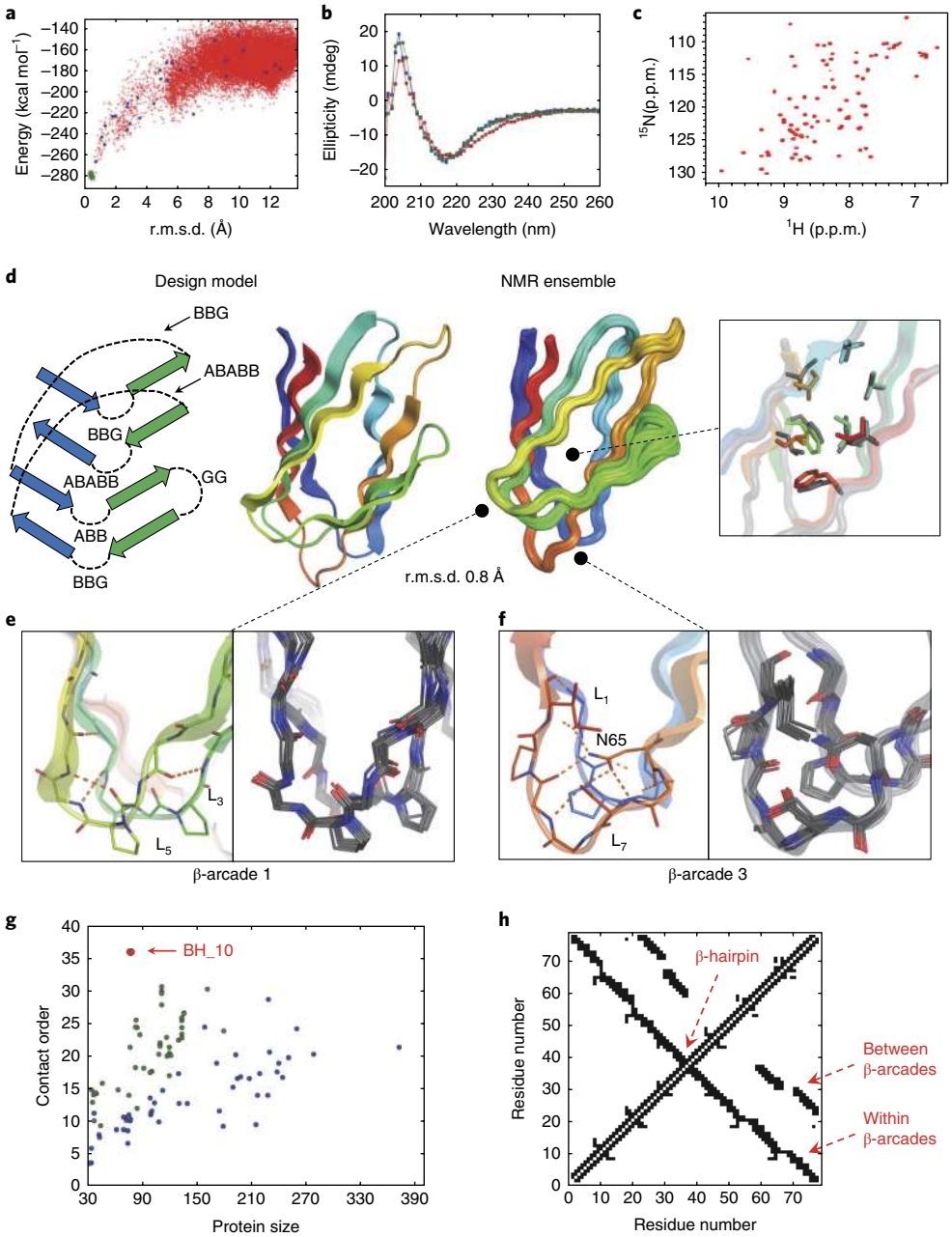


Fig. 3 | The NMR structure of BH_10 is nearly identical to the design model. a, Calculated BH_10 folding energy landscape. Each dot represents the lowest-energy structure obtained from ab initio folding trajectories starting from an extended chain (red dots), biased forward folding trajectories (blue dots) or local relaxation of the designed structure (green dots); the *x* axis is the Cα r.m.s.d. from the designed model and the *y* axis is the Rosetta all-atom energy. **b**, Far-ultraviolet CD spectra (blue, 25 °C; red, 95 °C; green, 25 °C after cooling). **c**, ^{1}H-^{15}N HSQC spectra obtained at 37 °C at a ^{1}H field of 800 MHz. **d**, NMR structure in comparison with the design model. Comparison of core side chain rotamers (NMR structure in gray and design in rainbow) (inset). Topology diagram showing the ABEGO torsion bins of all loop connections (left). Atomic coordinates for the design model are in Supplementary Dataset 1. **e**, Backbone hydrogen bonding of β-arcade 1 is well preserved across the NMR ensemble. **f**, Side chain interactions of N65 with backbone and side chains form a hydrogen-bonded network in β-arcade 3 that is well recapitulated in the NMR ensemble. **g**, Contact order of de novo–designed protein domains confirmed by high-resolution structure determination; all-α (blue), αβ (green) and all-β (red). The BH_10 design stands out with a contact order of 35.8 for a chain length of 78 residues. The domains are listed in Supplementary Tables 4 and 5. **h**, Contact map illustrating the large sequence separation of the contacts present in the BH_10 topology.

Contact order and sequence determinants of the BH_10 fold. The non-local character of BH_10 is of particular note: a large fraction of the contacting residues are distant along the linear sequence, with extensive strand pairing between the N- and C-terminal β-strands. The contact order of the structure (that is, the average separation along the linear sequence of residues in contact in the 3D structure) is higher than that for any previous single-domain protein designed de novo (Fig. 3g,h). Proteins with high contact order fold more slowly than those with low contact order as there is a greater loss in chain entropy for forming the first native interactions, and they tend to form long-lived non-native structures that can oligomerize or aggregate[31]. We have overcome the challenges in designing

Table 1 | NMR and refinement statistics for BH_10

	BH_10 (PDB 6E5C)
NMR distance and dihedral constraints	
Distance constraints	
Total NOE	659
Intraresidue	272
Inter-residue	387
Sequential ($\|i-j\|=1$)	222
Medium range ($2\leq\|i-j\|\leq4$)	33
Long range ($\|i-j\|\geq5$)	132
Intermolecular	0
Hydrogen bonds	0
Total dihedral angle restraints	156
φ	78
ψ	78
Structure statistics	
Violations (mean ± s.d.)	
Distance constraints (Å)[a]	0.30 ± 0.46
Dihedral angle constraints (°)[b]	9.30 ± 2.49
Max. dihedral angle violation (°)[b]	47.59
Max. distance constraint violation (Å)[a]	1.32
Deviations from idealized geometry	
Bond lengths (Å)	0.00 ± 0.00
Bond angles (°)	0.00 ± 0.00
Impropers (°)	0.00 ± 0.00
Average pairwise r.m.s.d. (Å)[c]	
Heavy	0.61 ± 0.13
Backbone	0.51 ± 0.11

[a]Distance constraint violations in the structural ensemble were calculated using a 7-Å universal upper distance bound for the NOE restraints assigned by AutoNOE-Rosetta. [b]Dihedral angle restraints were derived from TALOS-N. The violations were calculated for the core secondary structural regions of the ten lowest-energy models using a 15° cutoff beyond TALOS-N-predicted dihedral angles. [c]Pairwise r.m.s.d. was calculated among ten refined models for a core secondary structural region defined by residues 2–8, 11–18, 21–28, 32–36, 39–43, 46–53, 59–65 and 71–75.

non-local structures by focusing on backbones lacking internal strain and having maximal internal coherence, and programming β-strand orientation with highly structured loops.

One of the challenges in achieving high contact order through β-arches is to disfavor formation of more sequence-local β-hairpins. To evaluate in silico how each of our design features contributes to favoring β-arches over β-hairpins, we generated folding energy landscapes for a series of mutants of BH_10 in which loop hydrogen bonding, side chain packing of loop neighbors and loop local geometry were disrupted one at a time. For all conformations generated, we classified all the β-strand connections as β-arch or β-hairpin depending on strand-pairing formation, and calculated the overall frequency of β-hairpin formation for each pair of consecutive β-strands. As shown in Supplementary Fig. 9, disruption of packing within or between β-arch loops, removal of side chain–backbone hydrogen-bonding interactions and reducing loop geometry encoding by eliminating prolines all increase sampling of competing β-hairpin conformations, and thus substantially decrease sampling of β-arches and the target designed structure.

Discussion

The design of all-β globular proteins from first principles has remained elusive for two decades of protein design research. We have successfully designed a double-stranded β-helix de novo, as confirmed by the NMR structure of the design BH_10, based on a series of rules describing the geometry of β-arch loops and their interactions in more complex β-arcades. Our work also achieves two related milestones: the first accurate design of an all-β globular protein with exposed β-sheet edges, and the most non-local structure yet designed from scratch. Comparison of successful and failed designs suggests that folding and stabilization of the monomeric structure (and implicitly disfavoring competing topologies with more local strand pairings) is bolstered by loops containing side chain–backbone and backbone-backbone hydrogen bonds together with well-packed mixed aliphatic/aromatic side chains in the protein core, inward-pointing polar amino acids at strand edges and salt bridges between paired strands. Previous design studies on β-propellers[11] or parallel β-helices[12] have used naturally occurring backbone structures and consensus sequence information on the target fold families; this approach, while powerful, sheds less light on the key principles underlying β-sheet structure construction and does not allow the programming of new backbone geometries. The β-helix fold designed here is well suited for incorporating metal, ligand-binding and active sites, as illustrated by the broad functional diversity of cupin protein domains, which are the closest naturally occurring structural analogs. With the basic design principles now understood, our de novo design strategy should enable the construction of a wide range of β-helix structures tailored to a broad range of target ligands.

Initial advances in protein design were from algorithms that allowed rapid identification of a very-low-energy sequence for a given backbone structure. In recent years, progress has come from the realization that the requirements of burying hydrophobic residues in a core away from solvent while avoiding the burial of backbone polar groups without compensating hydrogen bonds, together with torsional restrictions on the peptide backbone, considerably constrain overall globular protein backbone geometry, particularly for β-sheet-containing proteins: it is much harder than originally expected to construct new backbones that have these properties. The de novo design of β-sheet-containing proteins advanced considerably following the elucidation of β-sheet design principles for construction of backbones meeting the above constraints while having desired geometries: for example, principles for controlling the chirality of β-hairpins[5], reducing strain in β-strands with glycine kinks[10], and combining β-bulges and register shifts to curve β-sheets[21]. The design rules described here are a considerable further advance as they provide control over β-arch connections between distinct β-sheets, and should enable the design of a broad range of β-protein families beyond the β-barrel and β-helix, with considerable medical and biotechnological potential; for example, the immunoglobulin fold widely utilized for binding and loop scaffolding in nature is topologically very similar to the double-stranded β-helices designed here, with a larger proportion of β-hairpins over β-arches.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41594-018-0141-6.

Received: 16 July 2018; Accepted: 11 September 2018;
Published online: 29 October 2018

References

1. Kortemme, T., Ramírez-Alvarado, M. & Serrano, L. Design of a 20-amino acid, three-stranded β-sheet protein. *Science* **281**, 253–256 (1998).
2. Searle, M. S. & Ciani, B. Design of β-sheet systems for understanding the thermodynamics and kinetics of protein folding. *Curr. Opin. Struct. Biol.* **14**, 458–464 (2004).

3. Hughes, R. M. & Waters, M. L. Model systems for β-hairpins and β-sheets. *Curr. Opin. Struct. Biol.* **16**, 514–524 (2006).
4. Marcos, E. & Adriano-Silva, D. Essentials of de novo protein design: methods and applications. *WIREs Comput. Mol. Sci.* **8**, e1374 (2018).
5. Koga, N. et al. Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012).
6. Hecht, M. H. De novo design of β-sheet proteins. *Proc. Natl Acad. Sci. USA* **91**, 8729–8730 (1994).
7. Plaxco, K. W., Simons, K. T. & Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994 (1998).
8. Quinn, T. P., Tweedy, N. B., Williams, R. W., Richardson, J. S. & Richardson, D. C. Betadoublet: de novo design, synthesis, and characterization of a β-sandwich protein. *Proc. Natl Acad. Sci. USA* **91**, 8747–8751 (1994).
9. Nanda, V. et al. De novo design of a redox-active minimal rubredoxin mimic. *J. Am. Chem. Soc.* **127**, 5804–5805 (2005).
10. Dou, J. et al. De novo design of a fluorescence-activating β-barrel. *Nature* **561**, 485–491 (2018).
11. Voet, A. R. D. et al. Computational design of a self-assembling symmetrical β-propeller protein. *Proc. Natl Acad. Sci. USA* **111**, 15102–15107 (2014).
12. MacDonald, J. T. Synthetic β-solenoid proteins with the fragment-free computational design of a β-hairpin extension. *Proc. Natl Acad. Sci. USA* **113**, 10346–10351 (2016).
13. Ottesen, J. J. & Imperiali, B. Design of a discretely folded mini-protein motif with predominantly β-structure. *Nat. Struct. Biol.* **8**, 535–539 (2001).
14. Hu, X., Wang, H., Ke, H. & Kuhlman, B. Computer-based redesign of β sandwich protein suggests that extensive negative design is not required for de novo β sheet design. *Structure* **16**, 1799–1805 (2008).
15. Hennetin, J., Jullian, B., Steven, A. C. & Kajava, A. V. Standard conformations of beta-arches in β-solenoid proteins. *J. Mol. Biol.* **358**, 1094–1105 (2006).
16. Lin, Y.-R. et al. Control over overall shape and size in de novo designed proteins. *Proc. Natl Acad. Sci. USA* **112**, E5478–E5485 (2015).
17. Kajava, A. V., Baxa, U. & Steven, A. C. β arcades: recurring motifs in naturally occurring and disease-related amyloid fibrils. *FASEB J.* **24**, 1311–1319 (2010).
18. Kuhlman, B. & Baker, D. Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA* **97**, 10383–10388 (2000).
19. Kuhlman, B. et al. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).
20. Richardson, J. S. & Richardson, D. C. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl Acad. Sci. USA* **99**, 2754–2759 (2002).
21. Marcos, E. et al. Principles for designing proteins with cavities formed by curved β sheets. *Science* **355**, 201–206 (2017).
22. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol.* **383**, 66–93 (2004).
23. Bradley, P. Toward high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868–1871 (2005).
24. Kuhn, M., Meiler, J. & Baker, D. Strand-loop-strand motifs: prediction of hairpins and diverging turns in proteins. *Proteins* **54**, 282–288 (2004).
25. Bradley, P. & Baker, D. Improved β-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation. *Proteins* **65**, 922–929 (2006).
26. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
27. Camacho, C. et al. BLAST: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
28. Evangelidis, T. et al. Automated NMR resonance assignments and structure determination using a minimal set of 4D spectra. *Nat. Commun.* **9**, 384 (2018).
29. Holm, L. & Laakso, L. M. Dali server update. *Nucleic Acids Res.* **44**, W351–W355 (2016).
30. Zimmermann, L. et al. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.* **430**, 2237–2243 (2018).
31. Clark, P. Protein folding in the cell: reshaping the folding funnel. *Trends Biochem. Sci.* **29**, 527–534 (2004).

Acknowledgements

We thank S. Rettie for mass spectrometry assistance and the rest of the Baker laboratory members for discussion. We acknowledge computing resources provided by the Hyak supercomputer system funded by the STF at the University of Washington, and Rosetta@Home volunteers in ab initio structure prediction calculations. Work carried out at the Baker laboratory was supported by the Howard Hughes Medical Institute, Open Philanthropy, and the Defense Threat Reduction Agency. E.M. was supported by a Marie Curie International Outgoing Fellowship (FP7-PEOPLE-2011-IOF 298976). G.O. was supported by a Marie Curie International Outgoing Fellowship (FP7-PEOPLE-2012-IOF 332094). This research was financially supported by the Ministry of Education, Youths and Sports of the Czech Republic within the CEITEC 2020 (LQ1601) project, the Grant Agency of Masaryk University (to K.T.), an R35 Outstanding Investigator Award to N.G.S. through NIGMS (1R35GM125034-01), and the Office of the Director, NIH, under High End Instrumentation Grant S10OD018455, which funded the 800-MHz NMR spectrometer at UCSC. IRB Barcelona is the recipient of a Severo Ochoa Award of Excellence from the Ministry of Economy, Industry and Competitiveness (government of Spain).

Author contributions

E.M. designed the research, carried out the loop structural analysis, set up the design method and performed design calculations. T.M.C. carried out design calculations, protein expression, purification and CD experiments. A.C.M. collected 4D NMR data. T.E. performed 4D-CHAINS analysis. S.N. carried out AutoNOE-Rosetta calculations. L.C. expressed isotopically labeled proteins and performed SEC-MALS analysis. L.G.N. designed the research and carried out design calculations. A.D. and G.O. helped in protein expression and characterization. K.T. and N.G.S. supervised NMR structure determination. D.B. designed and supervised the research. E.M., and D.B. prepared the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41594-018-0141-6.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to E.M. or D.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2018

Methods

Loop analysis. Loop connections between β-strands were collected from a non-redundant database of PDB structures obtained from the PISCES server[32] with sequence identity <30% and resolution ≤2 Å. We discarded those loops connecting β-strands with hydrogen-bonded pairing (β-hairpins), and the remaining 5,061 β-arch loops were subsequently analyzed. The ABEGO torsion bins of each residue position were assigned based on the definition shown in Supplementary Fig. 1, and the side chain directionality pattern of neighboring residues was defined according to Fig. 1a. The secondary structure of all residue positions was assigned with DSSP[33], and the last β-strand residue preceding and the first β-strand residue following the β-arch loop were chosen as the critical neighboring residues determining the side chain pattern of the loop. The loop bending was defined as the angle between the loop center of mass and the two strand positions adjacent to the loop. Those loops with bending angles larger than 120° were discarded from the analysis to correctly identify those loops producing a substantial change in the direction of the two connected β-strands. The loop dataset is available in Supplementary Dataset 2.

Backbone generation. We used the Blueprint Builder mover[5] of RosettaScripts[34] to build protein backbones by Monte Carlo fragment assembly using nine- and three-residue fragments compatible with the target secondary structure and torsion bins (ABEGO), as specified in the blueprints of every target topology. We used a polyvaline centroid representation of the protein and a scoring function accounting for backbone hydrogen bonding, van der Waals interactions (namely to avoid steric clashes), planarity of the peptide bond (omega score term), and compactness of structures (radius of gyration). Thousands of independent folding trajectories were performed and subsequently filtered. Owing to the non-local character of β-sheet contacts, we used distance and angle constraints to favor the correct hydrogen-bonded pairing between β-strand main chain atoms. For every target topology, we automatically set all pairs of residues involved in β-strand pairing to generate all constraints for backbone building. Protein backbones were filtered based on their match with the blueprint specifications (secondary structure, torsion bins and strand pairing), and subsequently ranked based on backbone hydrogen-bonding energy (lr_hb score term) and the total energy obtained from one round of all-atom flexible-sequence design (see next section).

Flexible-sequence design. Generated protein backbones were subjected to flexible-sequence design calculations with RosettaDesign[18,19] using the Rosetta all-atom energy function Talaris2014[35] to favor amino acid identities and side-chain conformations with low energy and tight packing. We performed cycles of fixed backbone design followed by backbone relaxation using the FastDesign mover[36] of RosettaScripts[34]. Designed sequences were filtered based on total energy, side-chain packing (measured with RosettaHoles[37], packstat and core side-chain average degree[21]), side chain–backbone hydrogen bond energy, and secondary structure prediction (match between the designed secondary structure and that predicted by PSPIRED[38] based on the designed sequence). Amino acid identities were restricted based on the solvent accessibility of protein positions, ensuring that hydrophobic amino acids were located in the core and polar ones in the surface. Further restrictions were imposed to improve sequence-structure compatibility in loop regions. Sequence profiles were obtained for naturally occurring loops with the same ABEGO string sequence, as previously[21].

For those blueprints that yielded the lowest energy designs we performed a second round with ten times more backbone samples. Backbones generated in this second round were subjected to more exhaustive sequence design by running multiple Generic Monte Carlo trajectories optimizing total energy and side chain average degree simultaneously, and then applied all filters described above.

Design of disulfide bonds and helix capping domain. We used the Disulfidize mover of RosettaScripts[34] to identify pairs of residue positions able to form disulfide bonds with a good scoring geometry. We searched for disulfide bonds between residues distant in primary sequence and with a disulfide score <−1.0. We designed a C-terminal helix capping domain (followed with a β-strand pairing with the first β-strand) using the backbone-generation protocol described above but starting from design BH_6. The structure of BH_6 was kept fixed during fragment assembly and the C-terminal domain was generated. Then, sequence design was performed for the C-terminal domain and those neighboring residues within 10 Å.

Sequence–structure compatibility. For assessing the local compatibility between designed sequences and structures we picked 200 naturally occurring fragments (9- and 3-mers) with sequences similar to the design, and evaluated the structural similarity (by r.m.s.d.) between the ensemble of picked fragments and the local designed structure. Those with overall low r.m.s.d. fragments, and therefore with high fragment quality, were subsequently assessed by Rosetta folding simulations using the Rosetta energy function 'ref2015'[39]. First, biased forward folding simulations[21] (using the three lowest r.m.s.d. fragments and 40 folding trajectories) were used to quickly identify those designs more likely to have funnel-shaped energy landscapes. Those designs achieving near-native sampling (r.m.s.d. to target structure below 1.5 Å) were then assessed by standard Rosetta ab initio structure prediction[22,23].

To evaluate the amount of β-hairpin sampling in each loop connection during ab initio structure prediction, we first detected all strand pairings formed in each generated decoy and then mapped the residues involved in those strand pairings to the secondary structure elements of the designed structure. After secondary structure mapping, pairings between strands consecutive in the sequence were counted as β-hairpins. The total count of β-hairpins sampled in each loop over the total number of generated decoys is a relative quantity of hairpin sampling allowed us to compare the β-hairpin propensity of different loops and mutants (see Supplementary Fig. 9).

Contact order. To evaluate the non-local character of protein structures we computed 'contact order' as the average sequence separation between pairs of Cα atoms within a distance of 8 Å and with a sequence separation of three residues at least.

Protein expression and purification. Genes encoding the designed sequences were obtained from Genscript and cloned into the pET-28b+ (with N-terminal 6× His tag and a thrombin cleavage site) expression vectors. Plasmids were transformed into *Escherichia coli* BL21 Star (DE3) competent cells, and starter cultures were grown at 37 °C in LB medium overnight with kanamycin. Overnight cultures were used to inoculate 500 ml LB medium supplemented with antibiotic, and cells were grown at 37 °C and 225 r.p.m. until an optical density (OD_{600}) of 0.5–0.7 was reached. Protein expression was induced with 1 mM IPTG at 18 °C and, after overnight expression, cells were collected by centrifugation (at 4 °C and 4,400 r.p.m. for 10 min) and resuspended in 25 ml of lysis buffer (20 mM imidazole and PBS). Resuspended cells were lysed in the presence of lysozyme, DNase and protease inhibitors. Lysates were centrifuged at 4 °C and 18,000 r.p.m. for 30 min, and the supernatant was loaded onto a nickel affinity gravity column pre-equilibrated in lysis buffer. The column was washed with three column volumes of PBS + 30 mM imidazole and the purified protein was eluted with three column volumes of PBS + 250 mM imidazole. The eluted protein solution was dialyzed against PBS buffer overnight. The expression of purified proteins was assessed by SDS–PAGE and mass spectrometry, and protein concentrations were determined from the absorbance at 280 nm measured on a NanoDrop spectrophotometer (Thermo Scientific) with extinction coefficients predicted from the amino acid sequences using the ProtParam tool (https://web.expasy.org/protparam/). Proteins were further purified by fast protein liquid chromatography size-exclusion chromatography using a Superdex 75 10/300 GL (GE Healthcare) column.

Circular dichroism. Far-ultraviolet CD measurements were carried out with the AVIV 420 spectrometer. Wavelength scans were measured from 260 to 195 nm at temperatures between 25 and 95 °C, using a 1-mm path-length cuvette. Protein samples were prepared in PBS buffer (pH 7.4) at a concentration of 0.2–0.4 mg ml^{-1}.

Size-exclusion chromatography combined with multiple-angle light scattering. SEC-MALS experiments were performed using a Superdex 75 10/300 GL (GE Healthcare) column combined with a miniDAWN TREOS multi-angle static light scattering detector and an Optilab T-rEX refractometer (Wyatt Technology). One-hundred-microliter protein samples of 1–3 mg ml^{-1} were injected onto the column equilibrated with PBS (pH 7.4) or TBS (pH 8.0) buffer at a flow rate of 0.5 ml min^{-1}. The collected data were analyzed with ASTRA software (Wyatt Technology) to estimate the molecular weight of the eluted species.

Protein expression of isotopically labeled proteins for NMR structure determination. Plasmids were transformed using standard heat-shock transformation into the Lemo21 expression strain of *E. coli* (NEB) and plated onto a minimal M9 medium containing glucose and kanamycin to maintain tight control over expression. A single colony was selected, inoculated into 50 ml of LB containing 50 μg ml^{-1} of kanamycin and grown at 37 °C with shaking overnight. After approximately 18 h, the 50-ml starter culture was removed and 25 ml was used to inoculate 500 ml of Terrific Broth (TB) containing 50 μg ml^{-1} kanamycin and mixed mineral salts[40]. The TB culture was grown at 37 °C with shaking at 250 r.p.m. until the OD_{600} reached a value of 1.0. Then, the culture was removed and the cells were pelleted by centrifugation at 4,000 r.p.m. for 15 min. The TB broth was removed and the pelleted cells were resuspended gently with 50 ml of 20 mM $NaPO_4$, 150 mM NaCl, pH 7.5. The resuspended cells were transferred into minimal labeling media, containing ^{15}N-labeled ammonium chloride at 50 mM and ^{13}C-labeled glucose to 0.25% (w/v), as well as trace metals, 25 mM Na_2HPO_4, 25 mM KH_2PO_4 and 5 mM Na_2SO_4. The culture was returned to 37 °C at 250 r.p.m. for 1 h in order to replace unlabeled nitrogen and carbon with labeled nitrogen and carbon. After 1 h, IPTG was added to 1 mM, the temperature was reduced to 25 °C and the culture was allowed to express overnight. The following morning, the culture was removed and the cells were pelleted by centrifugation at 4,000 r.p.m. for 15 min. The cells were resuspended with 40 ml of lysis buffer (20 mM Tris, 250 mM NaCl, 0.25% CHAPS, pH 8) and lysed with a Microfluidics M110P Microfluidizer at 18,000 p.s.i. The lysed cells were clarified using centrifugation at 24,000*g* for 30 min. The labeled protein in the soluble fraction was purified using immobilized metal affinity chromatography using standard methods (QIagen Ni-NTA resin). The purified protein was then concentrated to 2 ml and purified by FPLC

size-exclusion chromatography using a Superdex 75 10/300 GL (GE Healthcare) column into 20 mM $NaPO_4$, 150 mM NaCl, pH 7.5. The efficiency of labeling was confirmed using mass spectrometry.

Heteronuclear single-quantum coherence spectra. The designed BH_10 (0.81 mM) and BH_11 (0.64 mM) were exhaustively buffer exchanged into NMR buffer (50 mM NaCl, 20 mM sodium phosphate, pH 6.5, 0.01% (v/v) NaN_3, 4 mM EDTA and 1 U Roche protease inhibitor cocktail) in 95% H_2O/5% D_2O. 2D ^{1}H-^{15}N HSQC experiments were acquired at 37 °C with four scans, acquisition times of 72 ms (^{15}N) in the indirect dimension and recycle delay of 2 s.

Chemical shift assignment. For chemical shift assignment of BH_10, a set of two non-uniformly sampled (NUS) 4D NMR experiments, a 4D HC(CC-TOCSY(CO))NH and 4D ^{13}C-^{15}N edited HMQC-NOESY-HSQC, were acquired at 800 MHz at 37 °C as previously described[28]. For the 4D HC(CC-TOCSY(CO))NH experiment, spectrum widths were set to 12,500 (acquisition dimension) × 2,100 (^{15}N) × 8,000 (^{13}C) × 7,300 (^{1}H) Hz and acquisition times in the indirect dimensions of 60 ms (^{15}N), 8 ms (^{13}C) and 8 ms (^{1}H) using 16 scans and a recycle delay of 1 s. Spectra were acquired with 2,000 hypercomplex NUS points distributed over the indirectly detected dimensions (0.38% sparsity). For the 4D ^{13}C-^{15}N edited HMQC-NOESY-HSQC, spectrum widths were set to 12,500 (acquisition dimension) × 1,000 (^{15}N) × 8,000 (^{13}C) × 10,000 (^{1}H) Hz, respectively, and acquisition times in the indirect dimensions of 38 ms (^{15}N), 10 ms (^{13}C) and 20 ms (^{1}H) using 8 scans, a recycle delay of 1 s and a NOESY mixing time of 120 ms. Spectra were acquired with 4,000 hypercomplex NUS points distributed over the indirectly detected dimensions (0.32% sparsity). 4D NUS spectra were processed in NMRPipe[41] using SMILE reconstruction[42] and analyzed using NMRFAM-SPARKY[43]. For every ^{1}H-^{15}N HSQC peak, the corresponding planes in 4D-HCNH TOCSY and 4D-HCNH NOESY spectra were inspected and peaks were picked manually. The 4D peak lists were used as input for the 4D-CHAINS algorithm[28] to obtain sequence-specific resonance assignments of backbone and side chain atoms automatically. The overall assignment completeness reached 92%. No aromatic resonances were assigned. 4D-CHAINS assignments together with the 4D-HCNH NOESY peak list were used in AutoNOE-Rosetta for structure determination.

NOE assignment and structure determination using AutoNOE-Rosetta. To determine the structural models of the BH_10 target protein, we used CS-Rosetta[44] that provides AutoNOE-Rosetta[45] and RASREC-Rosetta[46] protocols. AutoNOE-Rosetta is an iterative NOE assignment method that utilizes RASREC-Rosetta to model protein structures de novo. These methods make use of valuable information contained within NMR chemical shifts about secondary and tertiary structures, and dynamics of proteins, to model targets of interest accurately[44,47]. The primary aim of AutoNOE-Rosetta is to label proton atoms to the unassigned NOESY cross-peaks by mapping their resonance frequencies to the assigned chemical shift frequencies. The resulting assignments can be utilized to create NOE-based distance restraints that aid the structure calculation process. The method begins by creating an initial mapping between the assigned chemical shift list and the unassigned NOESY cross-peak list. This mapping produces ambiguous assignments due to possible noise in the NOESY spectra[48,49]. These assignments undergo evaluation and filtering. The evaluation criteria rely on the symmetry of cross-peaks, chemical shift compatibility, intermediate structural model compatibility (in the subsequent stages of the protocol) and the participation of any NOE in a network of NOEs (network anchoring)[50]. The cross-peaks are eliminated if they lie along the diagonal in the NOESY spectra or their contribution to some of the evaluation criteria (such as network anchor score) is insignificant. The intensities of high-scoring NOE peaks are calibrated to produce distance restraints. These restraints are used to calculate structures within the highly parallel RASREC-Rosetta, which performs fragment assembly[44] using Monte Carlo methods and additional optimized algorithms[46]. This process of assigning NOEs and calculating structures is carried out iteratively across eight distinct stages. The final stage retains well-converged structural models alongside generated NOE restraints used for their calculation.

The process of setting up AutoNOE-Rosetta calculations is highly automated and accessible via a Python interface within the toolbox available at the CS-Rosetta website (https://csrosetta.chemistry.ucsc.edu). Before setting up NOE assignment and structure calculation runs, (1) NMR chemical shifts and target sequences are used to predict secondary structure (rigid regions) and flexible end regions from TALOS-N[51], (2) the flexible end regions are trimmed from sequence and chemical shift files because they cause deterioration of the performance of structure determination methods by inducing large numbers of degrees of freedom, and (3) the predicted secondary structure, together with trimmed chemical shift files, is used to select 200 structural fragments of amino acid lengths of three and nine, for each position in the target sequence. On completion of the previous steps, AutoNOE-Rosetta calculations are set up with target sequence, structural fragments, chemical shifts and unassigned NOESY cross-peak lists. For the BH_10 target protein, we obtained NMR chemical shifts from 4D-CHAINS[28] and additionally utilized amide to aliphatic (HCNH) unassigned NOESY cross-peak lists. Thereafter, we performed four rounds of AutoNOE-Rosetta calculations, where each round was supplied with a different restraint weight (standard restraint weights of 5, 10, 25 and 50 were used). All the calculation runs were evaluated using a function that assesses the all-atom energies[39] of the structural models and their convergence. After selecting the best-scoring restraint weight run, the ten models that exhibited the lowest energy within this run were selected. Commands to set up the calculations were used exactly as provided in the Supplementary Methods of a previous work[28]. Molprobity[52] was used to compute Ramachandran statistics for the ten lowest-energy structural models (100% of residues in favored regions of Ramachandran space, and 99% in favored regions) and deviations from the ideal geometry (Table 1).

Salt bridges. We used ESBRI[53] to predict salt bridges in the ten lowest-energy structural models. Out of 19 salt bridges predicted using ESBRI, AutoNOE-Rosetta recovers four salt bridges in the form of NOE contacts on the surface of the BH_10 protein. To identify salt bridges, we examined the NOE restraints assigned by AutoNOE-Rosetta between negatively charged glutamic acid or aspartic acid and positively charged arginine or lysine. We further filtered the restraints based on the upper distance bound of 4 Å in the ten lowest-energy structures. From these filters, we found that the salt bridges recapitulated by the NOE assignment module between the residue pairs are (15,62), (23,78), (33,64) and (35,62).

Hydrophobic core of BH_10. Buried residues were selected from the ten lowest-energy structural models using a 10 $Å^2$ solvent-accessible surface area threshold. There are 18 buried residues that contribute up to 70% of the total NOEs assigned by AutoNOE-Rosetta, and two of them are aromatic residues (F34 and F73). While 4D-CHAINS does not assign chemical shifts of side chain groups of aromatic residues (specifically aromatic rings), it provides respective chemical shift assignments of backbone atoms (Cα, Hα, N, H) and the β-carbon and β-proton (Cβ, Hβ) atoms. AutoNOE-Rosetta assigned a total of nine NOEs for the aromatic residues in the hydrophobic core, and the placement of aromatic side chains was optimized by the Rosetta packer algorithm. On close examination of the BH_10 structures, we found that the geometry of the two aromatic side chains was constrained by neighboring residues with methyl groups placed based on NOEs, supporting the accuracy of the aromatic side chain placement.

Visualization of protein structures and image rendering. Images of protein structures were created with PyMOL[54].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. The Rosetta macromolecular modeling suite (http://www.rosettacommons.org/) is freely available to academic and non-commercial users. Scripts and protocols used in this article for generating protein backbone blueprints, performing Rosetta design calculations and analyzing protein structures are all available on GitHub (https://github.com/emarcos/beta_sheet).

Data availability

NMR chemical shifts and NOESY cross-peak lists used to determine structures of BH_10 have been deposited in the BMRB with accession code 30495. Coordinates of the ten lowest-energy structures and the restraint lists have been deposited in the wwPDB as PDB 6E5C. The design model of BH_10 is available as Supplementary Dataset 1, and the loop dataset used to analyze the side chain patterns of naturally occurring β-arches is available in Supplementary Dataset 2. Other data are available from the corresponding authors upon reasonable request.

References

32. Wang, G. & Dunbrack, R. L. Jr. PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589–1591 (2003).
33. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
34. Fleishman, S. J. et al. RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PLoS ONE* **6**, e20161 (2011).
35. O'Meara, M. J. et al. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *J. Chem. Theory Comput.* **11**, 609–622 (2015).
36. Bhardwaj, G. et al. Accurate de novo design of hyperstable constrained peptides. *Nature* **538**, 329–335 (2016).
37. Sheffler, W. & Baker, D. RosettaHoles2: a volumetric packing measure for protein structure refinement and validation. *Protein Sci.* **19**, 1991–1995 (2010).
38. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
39. Alford, R. F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
40. Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expres. Purif.* **41**, 207–234 (2005).
41. Delaglio, F. et al. NMRPipe: a –spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).

42. Ying, J., Delaglio, F., Torchia, D. A. & Bax, A. Sparse multidimensional iterative lineshape-enhanced (SMILE) reconstruction of both non-uniformly sampled and conventional NMR data. *J. Biomol. NMR* **68**, 101–118 (2017).
43. Lee, W., Tonelli, M. & Markley, J. L. Nmrfam-Sparky: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **31**, 1325–1327 (2015).
44. Nerli, S., McShan, A. C. & Sgourakis, N. G. Chemical shift-based methods in NMR structure determination. *Prog. Nucl. Mag. Res. Sp* **106-107**, 1–25 (2018).
45. Lange, O. F. Automatic NOESY assignment in CS-RASREC-Rosetta. *J. Biomol. NMR* **59**, 147–159 (2014).
46. Lange, O. F. & Baker, D. Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. *Proteins* **80**, 884–895 (2012).
47. Berjanskii, M. V. & Wishart, D. S. Unraveling the meaning of chemical shifts in protein NMR. *Biochim. Biophys. Acta* **1865**, 1564–1576 (2017).
48. Nilges, M. A calculation strategy for the structure determination of symmetric dimers by ^{1}H NMR. *Proteins* **17**, 297–309 (1993).
49. Nilges, M. Ambiguous distance data in the calculation of NMR structures. *Fold Des.* **2**, S53–S57 (1997).
50. Herrmann, T., Güntert, P. & Wüthrich, K. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.* **319**, 209–227 (2002).
51. Shen, Y. & Bax, A. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J. Biomol. NMR* **56**, 227–241 (2013).
52. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010).
53. Costantini, S., Colonna, G. & Facchiano, A. M. ESBRI: a web server for evaluating salt bridges in proteins. *Bioinformation* **3**, 137–138 (2008).
54. The PyMOL Molecular Graphics System, Version 1.7.2 (Schrödinger, LLC, 2016).

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-09651-7 OPEN

Dishevelled-3 conformation dynamics analyzed by FRET-based biosensors reveals a key role of casein kinase 1

Jakub Harnoš[1,11], Maria Consuelo Alonso Cañizal[2,3,4], Miroslav Jurásek[5,6], Jitender Kumar[5], Cornelia Holler[7,8], Alexandra Schambony[7,8], Kateřina Hanáková[5,6], Ondřej Bernatík[1], Zbyněk Zdráhal[5,6], Kristína Gömöryová[1], Tomáš Gybel'[1], Tomasz Witold Radaszkiewicz[1], Marek Kravec[1], Lukáš Trantírek[5,9], Jan Ryneš[5], Zankruti Dave[1], Ana Iris Fernández-Llamazares[10], Robert Vácha[5,6], Konstantinos Tripsianes[5], Carsten Hoffmann[2,3,4] & Vítězslav Bryja[1,9]

Dishevelled (DVL) is the key component of the Wnt signaling pathway. Currently, DVL conformational dynamics under native conditions is unknown. To overcome this limitation, we develop the Fluorescein Arsenical Hairpin Binder- (FlAsH-) based FRET in vivo approach to study DVL conformation in living cells. Using this single-cell FRET approach, we demonstrate that (i) Wnt ligands induce open DVL conformation, (ii) DVL variants that are predominantly open, show more even subcellular localization and more efficient membrane recruitment by Frizzled (FZD) and (iii) Casein kinase 1 ε (CK1ε) has a key regulatory function in DVL conformational dynamics. In silico modeling and in vitro biophysical methods explain how CK1ε-specific phosphorylation events control DVL conformations via modulation of the PDZ domain and its interaction with DVL C-terminus. In summary, our study describes an experimental tool for DVL conformational sampling in living cells and elucidates the essential regulatory role of CK1ε in DVL conformational dynamics.

[1] Department of Experimental Biology, Faculty of Science, Masaryk University, Brno 62500, Czech Republic. [2] Department of Pharmacology and Toxicology, University of Würzburg, Würzburg 97078, Germany. [3] Rudolf Virchow Center for Experimental Biomedicine, University of Würzburg, Würzburg 97078, Germany. [4] Institute for Molecular Cell Biology, CMB—Center for Molecular Biomedicine, University Hospital Jena, Friedrich Schiller University Jena, Jena 07745, Germany. [5] CEITEC—Central European Institute of Technology, Masaryk University, Brno 62500, Czech Republic. [6] National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Brno, 62500, Czech Republic. [7] Max Planck Institute for the Science of Light, Erlangen 91058, Germany. [8] Biology Department, Developmental Biology, Friedrich-Alexander University Erlangen-Nüremberg, Erlangen 91058, Germany. [9] Institute of Biophysics, Academy of Sciences of the Czech Republic, v.v.i., Brno 612 65, Czech Republic. [10] Pepscan Therapeutics B.V., Lelystad 8243, The Netherlands. [11] Present address: Department of Cell, Developmental & Regenerative Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. Correspondence and requests for materials should be addressed to V.B. (email: bryja@sci.muni.cz)

Wnt signaling pathway is one of the signaling pathways, whose dysfunction or deregulation is linked to developmental defects, inherited diseases, and many types of cancer[1]. Wnt proteins have been shown to activate several Wnt signaling pathways. The best known Wnt/β-catenin (canonical) pathway typically controls cell proliferation and differentiation by the regulation of transcription, whereas other (i.e. non-canonical) branches typically regulate cell polarity and migration by the reorganization of cytoskeleton[2–4]. Wnt signals are transduced via the membrane receptor Frizzled (FZD) and the intracellular protein Dishevelled (Dsh in *Drosophila*, DVL1–3 in humans). There is a general agreement based on strong genetic evidence in *Drosophila*[3] and mouse[5,6] that DVL protein(s) plays a crucial role in all major branches of Wnt signaling and serves as a signaling hub.

All DVL proteins have a conserved architecture consisting of three well-defined domains: an N-terminal DIX domain (Dishevelled and Axin), a central PDZ domain (Post-synaptic Density Protein-95, Disc Large Tumor Suppressor, and Zonula Occludens-1), and a C-terminal DEP domain (Dishevelled, Egl-10, and Pleckstrin). Crystal structures of all three isolated DVL domains are known[7–9]. The regions linking the three domains and the long C-terminus (200 aa) are intrinsically disordered[2].

DVL molecules can multimerize to form high-order structures called DVL signalosomes, which are crucial for the subsequent downstream activation of Wnt/β-catenin signaling[10–12]. Signalosome formation is mediated via head-to-tail multimerization of DVL's DIX domains[10] (and/or DIX domain of Axin protein[11–13]). Recently, dimerization via the DEP domain has also been reported and it was proposed that DEP–DEP interaction mediated by domain swapping cooperates with DIX-dependent DVL signalosome formation[14].

Several pieces of information suggest that DVL can also interact intramolecularly. Specifically, it has been shown in vitro that at least two regions located in the C-terminus of DVL— (i) a nuclear export signal and (ii) the terminal 7 aa in the C-terminus of DVL—can interact with the PDZ domain[15–17]. These reports suggested that DVL can exist in a closed and open conformation (s) undergoing multiple structural transitions that control DVL function. However, there is neither an experimental approach to study this phenomenon directly nor clear evidence that DVL exists in multiple conformations in vivo.

In this study we use an experimental system that enables the analysis of DVL conformational sampling in living mammalian cells. We design and characterize several biosensors of DVL3 conformation based on fluorescein arsenical hairpin binder (FlAsH)-based Förster resonance energy transfer (FRET). This single-cell FRET method allows site-selective detection of protein conformation dynamics in living cells and has already helped to decipher various biological problems[18–20]. FlAsH is a small organic compound, which recognizes and binds to CCPGCC tag[18]. The FlAsH–CCPGCC complex becomes fluorescent and can be used as an energy acceptor in the FRET pair with ECFP (enhanced cyan fluorescent protein). The main advantage of the CCPGCC tag is its size—only 6 aa, which reduces interference with the biological function of the protein of interest[18]. Using FlAsH FRET DVL3 sensors we describe the conformational dynamics of DVL3 in living cells upon Wnt pathway stimulation and the key function of Casein kinase 1 ε (CK1ε) in this process.

Results

Design and generation of the FRET-based DVL3 biosensors. FlAsH-based FRET represents a powerful technique that allows monitoring of the protein conformation in biological setups, where large tags can interfere with protein complex formation and function[18–20]. FlAsH molecule forms a fluorescent complex with a 6 aa long sequence (CCPGCC) that can be shut off by addition of BAL (British Anti-Lewisite or 2,3-dimercapto-1-propanol)[18]. The general scheme of how FlAsH-based FRET works is depicted in Fig. 1a and in Supplementary Fig. 1a. We designed and generated four different DVL3 sensors (FlAsH I–IV), which differ in the position of the CCPGCC tag. Each DVL3 sensor contained the ECFP tag at the N-terminus and one CCPGCC internal tag, located in an intrinsically disordered region (IDR) or at the C-terminus of DVL3. The CCPGCC tags were inserted at the most disordered region of each IDR predicted in silico using PONDR-FIT tool[21] as shown in Fig. 1b. Biological and signaling properties of all four DVL3 FlAsH sensors were indistinguishable from wild-type (wt) ECFP-DVL3 in the following assays (performed in HEK293 DVL1-2-3$^{-/-}$ cell line[22] in order to avoid interaction with endogenous full-length DVL isoforms): activation of the Wnt/β-catenin pathway monitored by Dual Luciferase TopFlash/Renilla reporter gene assay (Supplementary Fig. 1c); CK1ε-dependent DVL electrophoretic mobility shift assay (Supplementary Fig. 1d); and changes of DVL subcellular localization induced by CK1ε (Supplementary Fig. 1e). Data are summarized in Fig. 1c.

The ECFP-DVL3 FlAsH I, II, and III (but not IV) sensors showed robust basal intramolecular FRET measure after BAL addition (Fig. 1d, left panel), which suggests that the ECFP and CCPGCC tags are in physical proximity. Since it has been reported that DVL can undergo intermolecular multimerization via its DIX and DEP domains[10,14], intermolecular FRET among different DVL3 molecules can also occur. To assess the intermolecular FRET efficiency, we generated N-terminally HA-tagged DVL3 FlAsH I–IV constructs without the N-terminal ECFP tag. Despite the fact that the ECFP and CCPGCC tags (labeled by FlAsH molecule) showed almost complete co-localization in dots for all four sensors (Supplementary Fig. 1b), the intermolecular FRET efficiency was negligible, except FlAsH III (Fig. 1d, right panel). These results demonstrate that ECFP-DVL3 FlAsH I, II, and III sensors represent a useful tool for the analysis of DVL conformation in living cells that is not affected (FlAsH I, II) or only minimally affected (FlAsH III) by intermolecular FRET due to the DVL–DVL oligomerization.

CK1ε changes intramolecular FRET in DVL3 sensors. As a result of Wnt pathway activation, DVL is heavily phosphorylated[23–26]. It has been further shown that the major DVL kinase responsible for Wnt-induced phosphorylation of DVL[27–29] is CK1ε. In order to analyze whether and how CK1ε affects DVL3 conformation, we (i) used CK1δ/ε inhibitor PF670462 (refs. [29,30]) to pharmacologically inhibit endogenous CK1ε and (ii) promoted DVL3 phosphorylation by overexpression of exogenous CK1ε. Such manipulation of CK1ε activity dramatically affects DVL3 properties such as: (i) CK1ε-DVL3-mediated activation of the Wnt/β-catenin pathway monitored by Dual Luciferase TopFlash/Renilla reporter gene assay (Fig. 2a); (ii) phosphorylation status of DVL3 assessed either as a phosphorylation-dependent mobility shift on SDS-PAGE or as an increased signal for pS643-DVL3 phosphorylation-specific antibody—a well-defined[29] target site for CK1ε (Fig. 2b), and (iii) changes in the subcellular localization of DVL3 (ref. [31]) analyzed by immunofluorescence (Fig. 2c). These properties were analyzed for all four ECFP-DVL3 FlAsH sensors (Supplementary Fig. 2a) and were found indistinguishable from wt FLAG-DVL3 (compare Supplementary Fig. 2a and Fig. 2a–c). This result shows that neither ECFP nor CCPGCC tag addition in any of the four sensors affects the biological response of DVL3 to CK1ε.

To determine how CK1ε influences DVL3 conformation sampling, we analyzed basal FRET in ECFP-DVL3 FlAsH I-IV

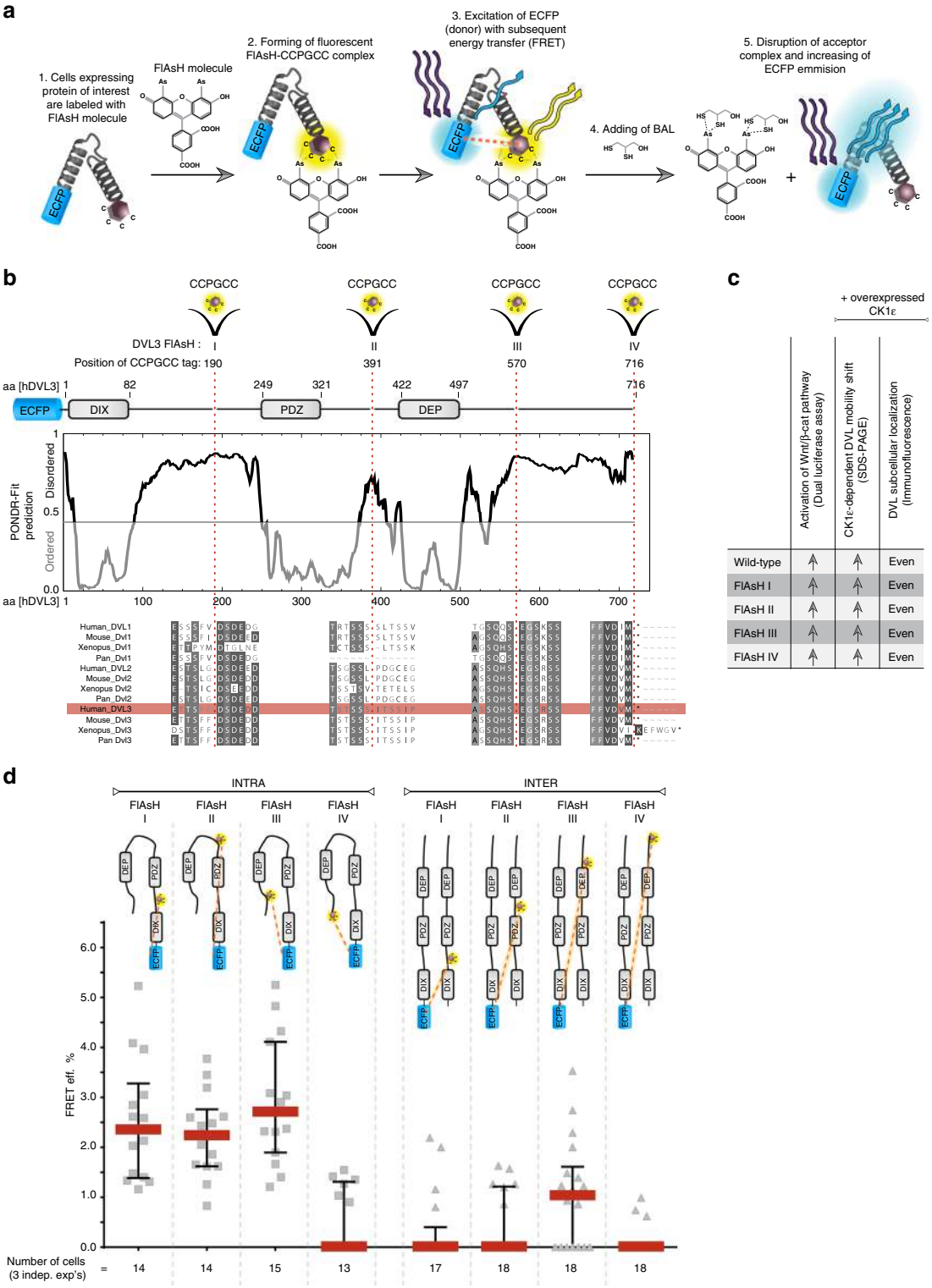


Fig. 1 The FlAsH-based FRET DVL3 sensors. **a** The general scheme of the FlAsH-based FRET in vivo approach (in further detail in Supplementary Fig. 1a). **b** The design of four DVL3 FlAsH I–IV sensors with the CCPGCC tag and N-terminal ECFP tag. The insertions of the CCPGCC tag were placed at the positions with highest disordered prediction scores (PONDR-Fit[21])—one CCPGCC tag per a linker region (FlAsH I, II, III) and the C-terminus (FlAsH IV). Multiple sequence alignment of the Dvl/DVL sequences at the site of insertion is shown below. Residues with >80% similarity are highlighted. Sequence of human DVL3 used for cloning as a template is shown in red box. **c** Biological properties of four ECFP-DVL3 FlAsH sensors are indistinguishable from wild-type ECFP-DVL3 (for details see Supplementary Fig. 1c–e). **d** The intramolecular and intermolecular FRET efficiency in the DVL3 FlAsH I–IV sensors in HEK293 wild-type cells. The position of the fluorophores in DVL3 molecules in both experimental setups are schematized above the graph. One data point corresponds to one analyzed cell. Data from three independent transfections were merged. Data in **d** represent median ± interquartile range. FRET eff., FRET efficiency; BAL, British anti-Lewisite; ECFP enhanced cyan fluorescent protein

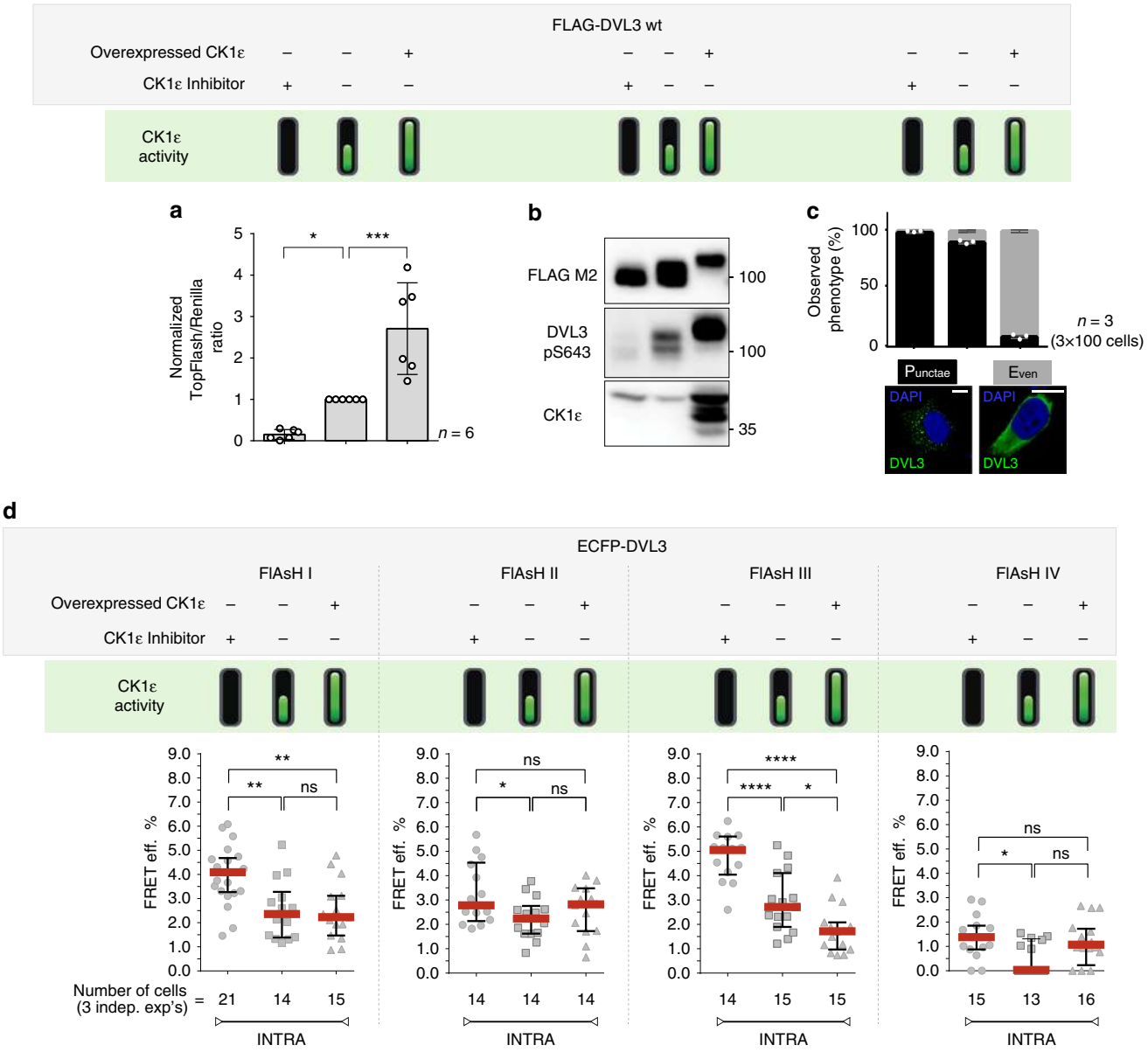


Fig. 2 DVL3 FlAsH I–IV sensors uncover the role of CK1ε in DVL3 conformation. **a** The effects of the CK1δ/ε inhibitor PF-670642 and the overexpressed CK1ε on the downstream Wnt/β-catenin signaling of FLAG-DVL3 was quantified by TopFlash assay, **b** phosphorylation status detected either as phosphorylation-dependent mobility shift on SDS-PAGE or as an increased signal for pS643-DVL3 phosphorylation-specific antibody, and **c** changes in the localization of DVL3 (punctae phenotype plotted as white dots); DAPI was used for nuclei staining; scale bars: 10 µm. **d** Measurements of the intramolecular FRET efficiency of the ECFP-DVL3 FlAsH sensors I–IV in HEK293 wild-type cells are shown. One data point corresponds to one analyzed cell; datapoints from three independent transfections were merged. Data in **a** and **c** represent mean ± S.D., data in **d** represent median ± interquartile range. Statistical significance in **a**/**d** was analyzed by one-way ANOVA test with Gaussian distribution and Tukey's post-test (*$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$, ****$p \leq 0.0001$; ns, not significant, $p > 0.05$)

sensors in: (i) cells treated with the CK1δ/ε inhibitor (PF670462, 10 µM), (ii) control cells possessing active endogenous CK1ε, and (iii) cells overexpressing CK1ε (Fig. 2d, for intermolecular FRET efficiency controls see Supplementary Fig. 2b). Changes in CK1ε levels and activity affected FRET efficiency to a different extent for each sensor. The most prominent differences were detected using the FlAsH III sensor, for which the CK1δ/ε inhibitor consistently increased the FRET efficiency, whereas the CK1ε overexpression decreased it. These experiments implied that CK1ε inhibition results in a more compact (i.e. closed) conformation of DVL3, while CK1ε overexpression leads to a looser (i.e. open) conformation. In order to dissect mechanistically how CK1ε affects the DVL conformations in vivo, we selected DVL3 FlAsH III sensor for a rigorous investigation.

Identification of DVL3 regions in direct contact with CK1ε. As a next step, we mapped in detail the interaction site(s) of DVL3 and CK1ε, in order to generate DVL3 variants incapable of CK1ε binding while keeping all structured domains intact. We hypothesized that IDRs significantly contribute to the interaction of DVL3 and CK1ε. The binding epitopes within IDRs are typically defined by linear peptide motifs[32] that can be rapidly analyzed in

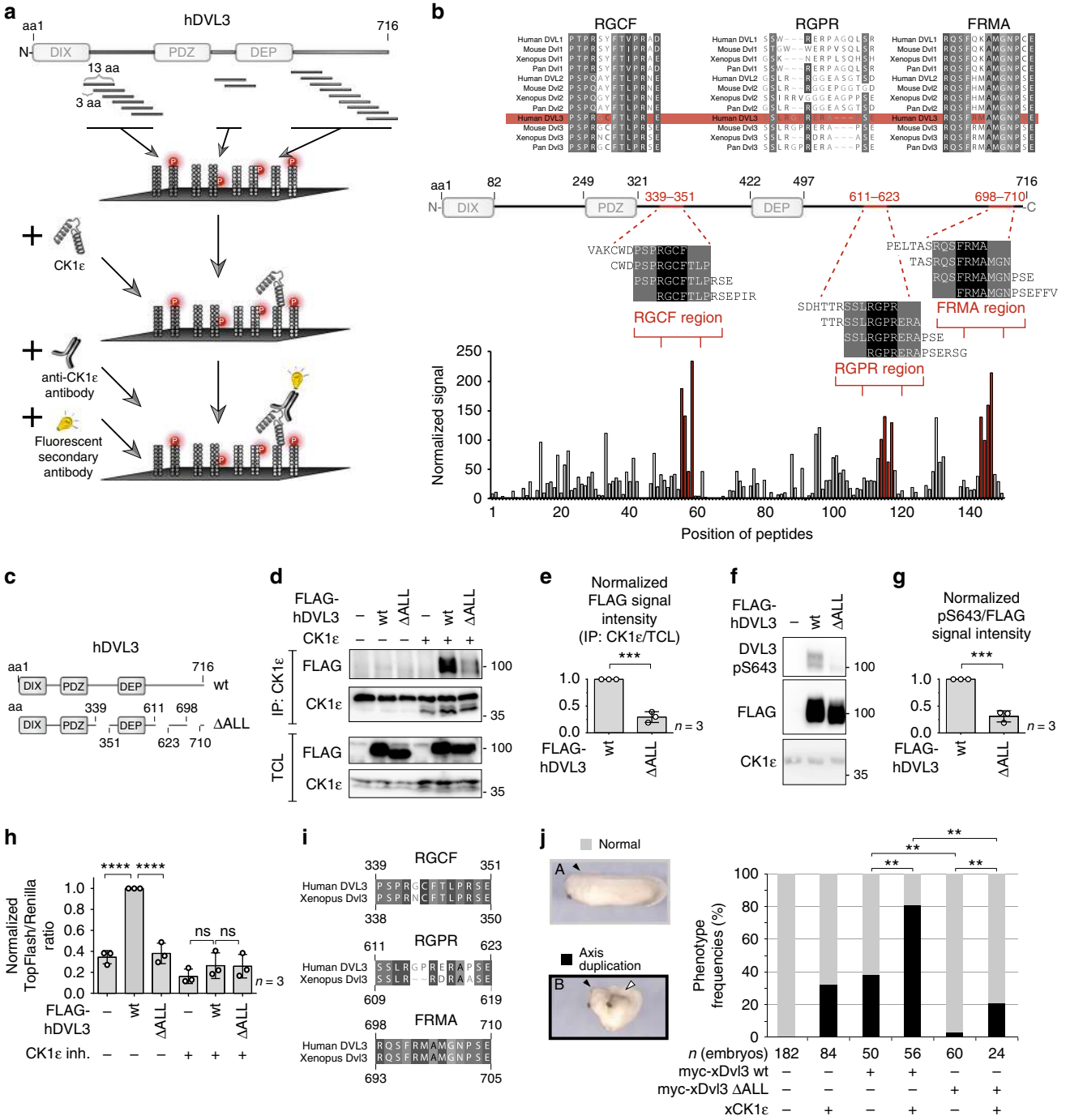


Fig. 3 The molecular and functional analysis of DVL3 regions interacting with CK1ε. **a** The general workflow of a peptide array analysis: immobilized peptides were incubated with CK1ε, then with anti-CK1ε antibody followed by fluorescent secondary antibodies, and detected by reading the fluorescence intensity. Peptide array contained the non-modified or phosphorylated peptide variants from the intrinsically disordered regions (IDRs) according to phosphorylation pattern by CK1ε, as mapped earlier[29] (see Supplementary Fig. 3). **b** Identification of three regions (named RGCF, RGPR, and FRMA regions by their central 4 aa sequences), which bind CK1ε with high affinity. Multiple sequence alignment for the RGCF, RGPR, and FRMA regions of various Dvl/DVL isoforms is shown above. Residues with >80% similarity are highlighted and human DVL3 sequence is denoted by red box; only non-modified (i.e. non-phosphorylated) peptides are shown in this graph. **c** Generation of the N-terminal FLAG-tagged DVL3 ΔALL variant lacking the interaction interfaces (RGCF, RGPR, FRMA regions) and its subsequent analysis by **d** coimmunoprecipitation and **e** its quantification, **f** by western blot detection of the pS643 phosphorylation level and **f** its quantification, **g** and by Topflash Reporter Assay for the downstream Wnt/β-catenin signaling. **h** The multiple sequence alignment of *Xenopus* Dvl3 and human DVL3 sequences in the RGCF, RGPR, and FRMA regions is shown. **i** Analysis of the activity of the ΔALL variant derived from *Xenopus* xDvl3 in the Wnt/β-catenin canonical signaling (in the *Xenopus laevis* embryos). **j** Left: Representative image of control (low or no activity of the Wnt/β-catenin pathway; in a gray box) or duplicated (high activity; in a black box) axis in the *Xenopus laevis* embryos. Right: Quantification of the *Xenopus laevis* embryos with wild-type xDvl3 and the ΔALL variant. Experiments in **d**–**f** were performed in HEK DVL1-2-3$^{-/-}$ cell line. Data in **e**, **g**, **h**, **j** represent mean ± S.D. Data in **h** and **j** were analyzed by one-way ANOVA test with Gaussian distribution; Tukey's post-test was used for statistical analysis (*, $p \le 0.05$; **, $p \le 0.01$; ***, $p \le 0.001$, ****, $p \le 0.0001$; ns, not significant, $p > 0.05$); data in **e** and **g** were analyzed by Student's *t*-test

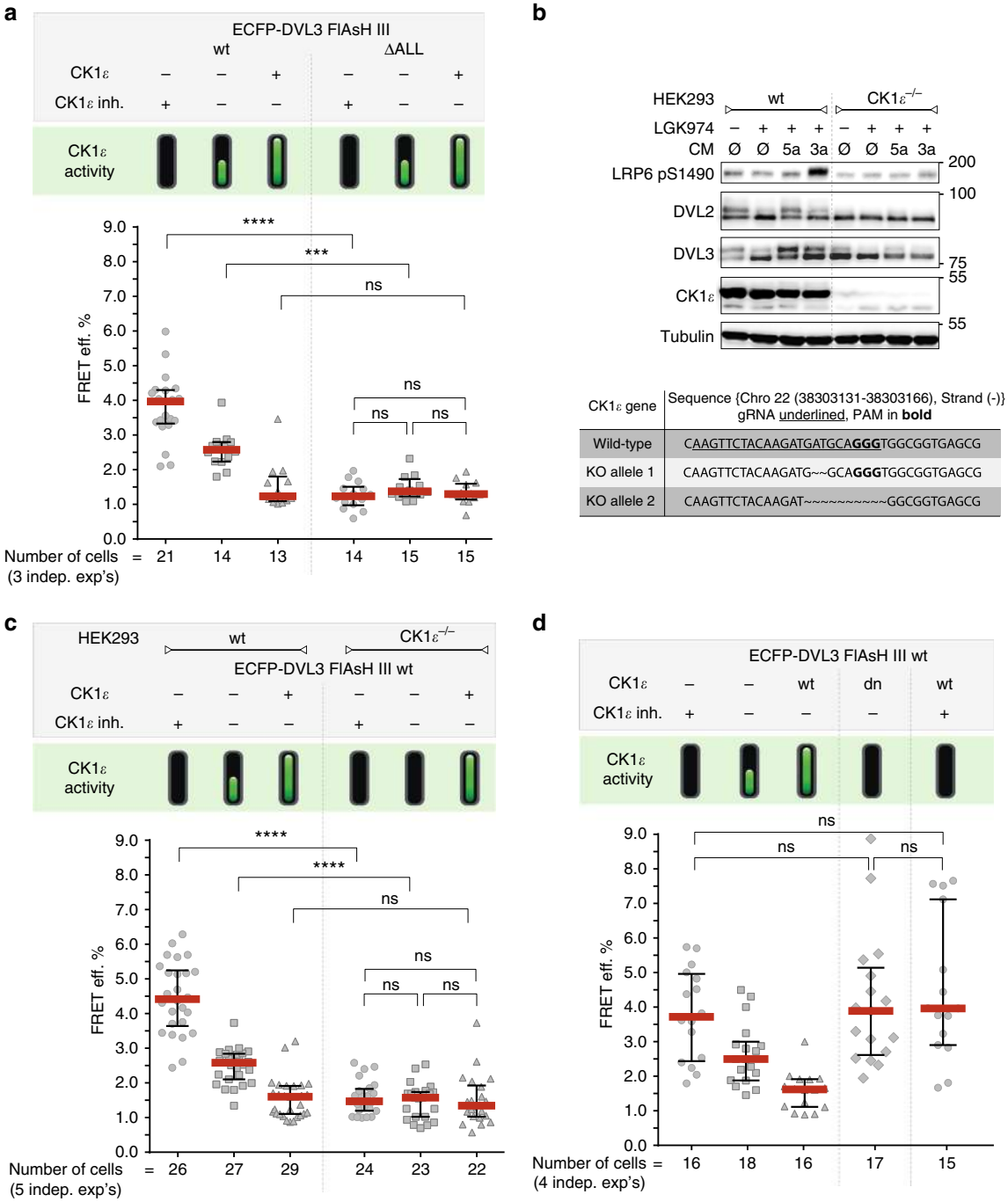


Fig. 4 CK1ε is required for the conformational dynamics of DVL3. **a** Sequences corresponding to the ΔALL variant were deleted in the ECFP-DVL3 FlAsH III sensor and the intramolecular FRET efficiency measurement of this variant (ECFP-DVL3 FlAsH III ΔALL sensor) in HEK293 wild-type cells is shown. **b** Generation of the CK1ε$^{-/-}$-deficient HEK293 cells using the CRISPR-Cas9 system and analysis of their capacity to respond to the Wnt-3a and Wnt-5a ligands was analyzed by western blotting. Bottom: Sequencing results for CK1ε locus targeted in CK1ε$^{-/-}$ cells are shown; sequences of gRNA, which were used, are underlined, the PAM sequence is in bold. **c** Measurements of the intramolecular FRET efficiency of the wt DVL3 FlAsH sensor III in HEK293 wild type and CK1ε$^{-/-}$-deficient cells. **d** Measurements of the intramolecular FRET efficiency of the wt DVL3 FlAsH III sensor in HEK293 wild-type cells with dominant negative (dn) variant of CK1ε and wt CK1ε treated with the CK1δ/ε inhibitor are shown. Data in **a**, **c** and **d**: one dot represents one cell; data from three to five independent transfections were merged. Median ± interquartile range is indicated; one-way ANOVA test with Gaussian distribution and Tukey's post-test was used for statistical analysis (*, $p \le 0.05$; **, $p \le 0.01$; ***, $p \le 0.001$, ****, $p \le 0.0001$; ns, not significant, $p > 0.05$). CM: conditional medium, Ø: control, 5a: Wnt-5a, 3a: Wnt-3a

peptide microarrays. For our purposes, we designed a peptide microarray (13-meric peptides overlapping by 10 residues) covering the IDRs of DVL3 (Fig. 3a).

The human DVL3 peptide library also contained phosphorylated versions of those peptides, where phosphorylation sites for CK1ε were previously identified[29] (Supplementary Fig. 3). After incubation with recombinant CK1ε, we discovered three peptide clusters—named RGCF (aa 339–351), RGPR (aa 611–623), and FRMA (aa 698–710)—as potential interaction sites with CK1ε (Fig. 3b). The MS/MS-based analysis of the wild-type DVL3

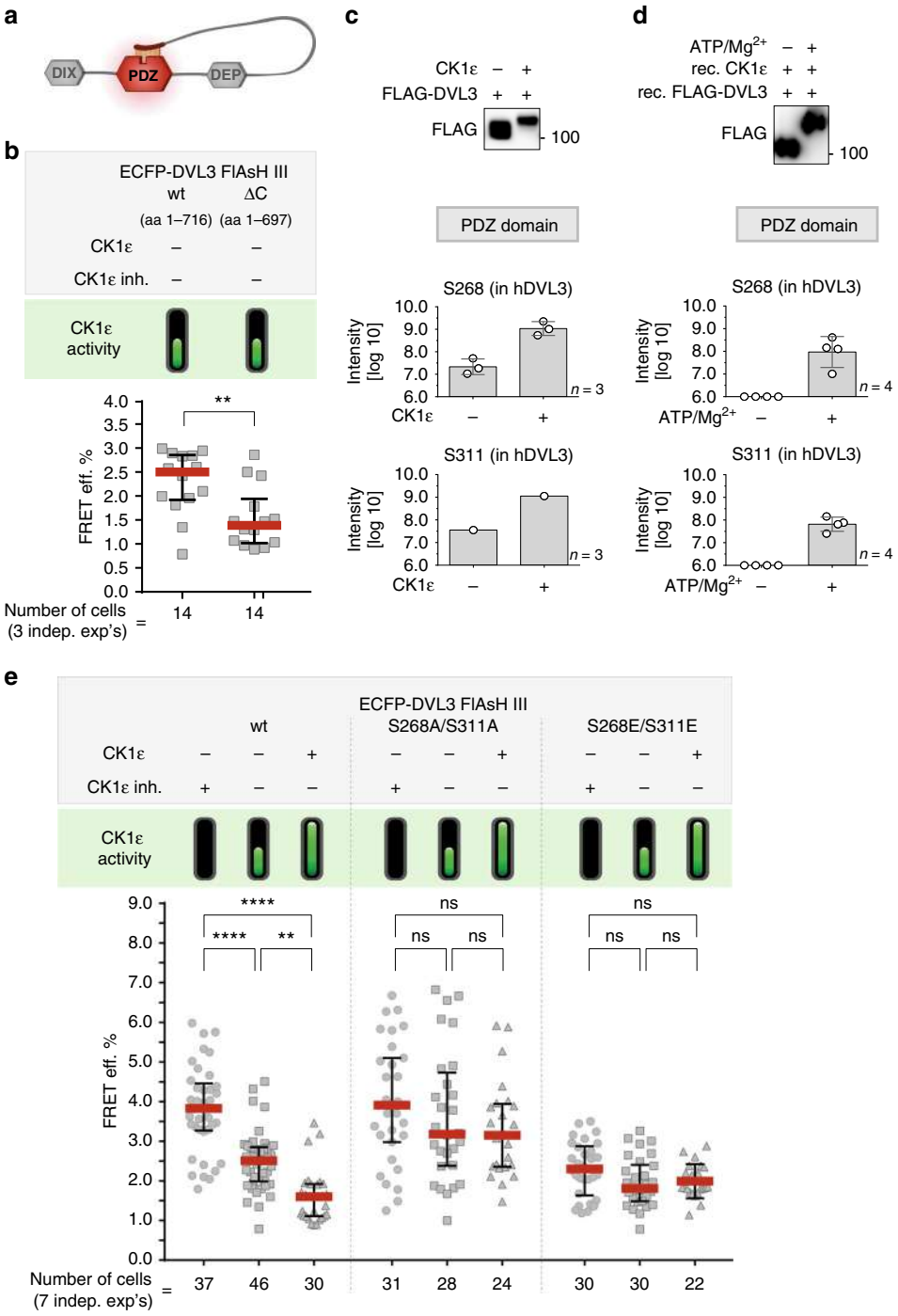


Fig. 5 Phosphorylation of PDZ domain controls the conformational dynamics of DVL3. **a** Schematic depiction of the closed conformation of DVL proposed here[16], where seven last C-terminal aa (sequence: EFFVDIM) interact with the peptide-binding pocket of the PDZ domain. **b** Measurements of the intramolecular FRET efficiency of the wt DVL3 FlAsH sensors III (aa 1–716 in human DVL3) and the ΔC variant (aa 1–697) in HEK293 wild-type cells. **c** The western blot-based (above) and MS/MS-based (below) analyses of the CK1ε-induced phosphorylation of serine residues present in the PDZ domain of human DVL3. FLAG-DVL3 wt was overexpressed with/without CK1ε wt in HEK293 wt cells, immunoprecipitated, and the level of phosphorylation was analyzed by MS/MS. CK1ε-induced phosphorylation of S268 and S311 in DVL3 PDZ domain was detected. **d** In vitro kinase assay with recombinant FLAG-DVL3 and CK1ε analyzed by western blot (above) and MS/MS (below) confirms that S268 and S311 are direct phosphorylation sites of CK1ε. Values in **c** and **d** show absolute intensity of phosphorylated peptides plotted on a log10 scale. The detection limit is approximately 1.10^6, i.e. 6.0. Individual datapoints represent biological replicates. **e** Comparison of the intramolecular FRET efficiency of wt DVL3 FlAsH III sensor, non-phosphorylatable variant (S268A/S311A), and phosphorylation-mimicking variant (S268E/S311E) in wt HEK293 cells is shown. One data point corresponds to one analyzed cell; datapoints from 3 (**b**) and 7 (**e**) independent transfections were merged. Data in **b** and **e** represent median ± interquartile range and data in **e** were analyzed by one-way ANOVA test with Gaussian distribution and Tukey's post-test (*, $p \leq 0.05$; **, $p \leq 0.01$; ***, $p \leq 0.001$, ****, $p \leq 0.0001$; ns, not significant, $p > 0.05$). Data in **b** were analyzed by Student's *t*-test. Data in **c** and **d** represent mean ± S.D.

overexpressed in HEK293 cells showed that serine residues in these regions (S350 in RGCF; S611, S612, and S622 in RGPR) were only negligibly phosphorylated, in contrast to S700 in FRMA that was found constitutively phosphorylated (Supplementary Fig. 4b). When CK1ε was overexpressed, all residues in RGCF and RGPR regions were phosphorylated (Supplementary Fig. 4c, d). Importantly, phosphorylated variants of these peptides showed negligible interaction with CK1ε in the peptide array (Supplementary Fig. 4e). The quantification of the in vitro binding between the peptides corresponding to these regions and recombinant CK1ε using fluorescence anisotropy (FA) showed affinities close to 1 μM for all three peptides but negligible interaction for their phosphorylated variants (Supplementary Fig. 4a). This suggests that the interaction is phosphorylation-sensitive and CK1ε is released once the binding epitopes get phosphorylated.

Deletion of these regions in DVL3 (hereafter referred as DVL3 ΔALL variant) (Fig. 3c) drastically reduced, but not completely abolished, the interaction with CK1ε as confirmed by coimmunoprecipitation and resulted in the diminished phosphorylation at the CK1ε target site S643 (Fig. 3d–g). Functionally, DVL3 ΔALL failed to activate the downstream Wnt/β-catenin pathway in the TopFlash reporter assay (Fig. 3h) in HEK293 DVL1-2-3$^{-/-}$ cells.

Intriguingly, the DVL3-CK1ε-binding sites show high sequence conservation (Fig. 3b, top) and are almost identical in humans and *Xenopus* (Fig. 3i). This allowed us to analyze the functional consequences of these deletions also in vivo. The activation of the Wnt/β-catenin pathway results in the axis duplication in *Xenopus*, which is a well-defined functional Wnt/β-catenin signaling readout (Fig. 3j, left). First, we injected a dose of mRNA encoding Dvl3 wild type into the marginal zone of the ventral blastomeres of the four-cell stage *Xenopus laevis* embryos to induce double axis formation (Fig. 3j, right). Not surprisingly, the xDvl3 ΔALL variant (lacking aa 338–350, 609–619, and 693–705 in xDvl3) showed dramatically reduced capacity to induce axis duplication both in the presence and absence of exogenous xCK1ε (Fig. 3j, right). Taken together, these data demonstrate that the identified DVL3 regions represent evolutionary conserved bona fide interaction sites for CK1ε, whose deletion abolishes both CK1ε binding and CK1ε-dependent functions of DVL3.

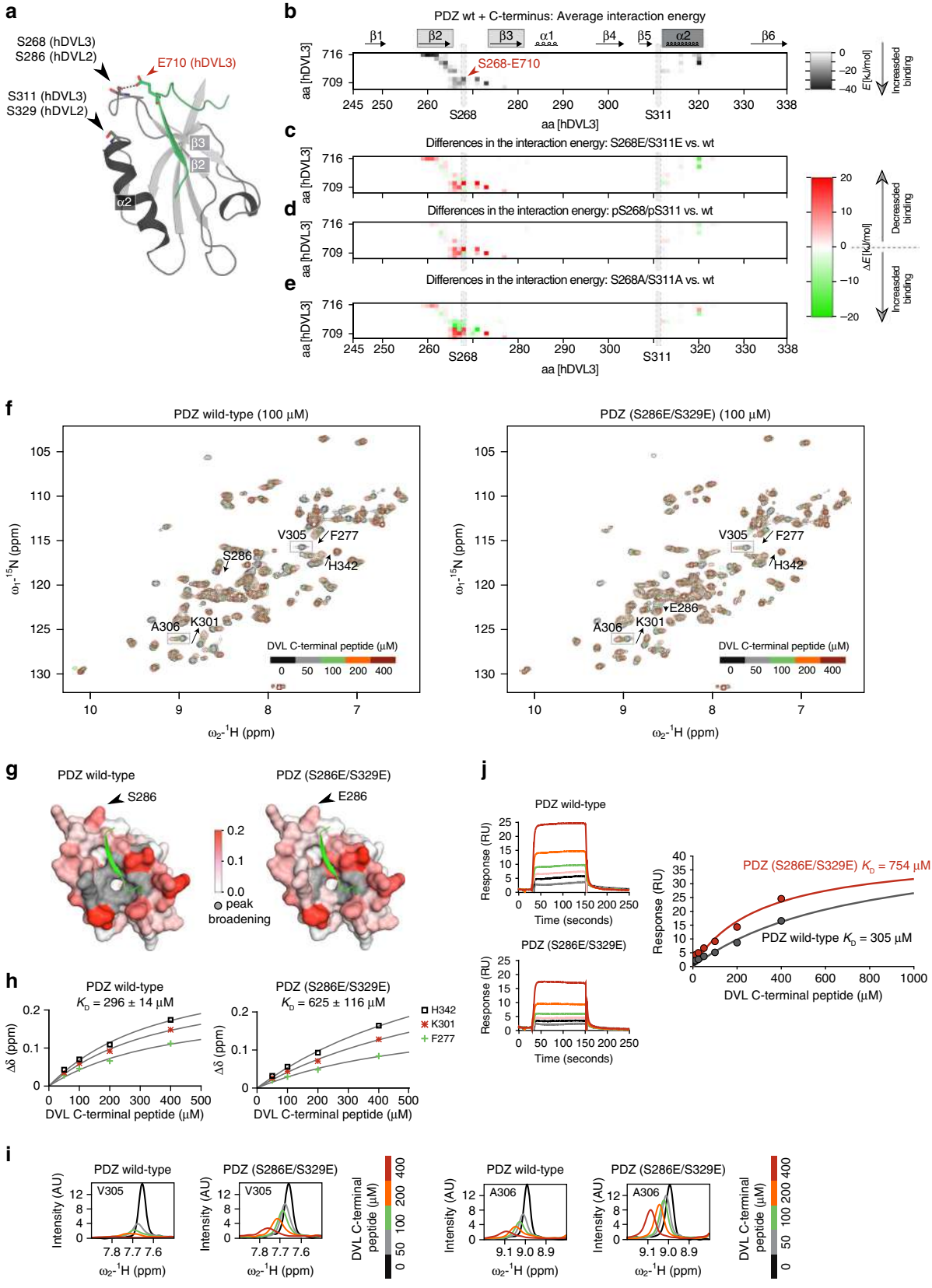
CK1ε is required for the conformational dynamics of DVL3. As the DVL3 ΔALL variant is incapable of complete interactions with CK1ε, we further examined the role of CK1ε in the conformational dynamics of DVL3. Using the FlAsH III sensor as a template, we generated and analyzed the ECFP-DVL3 FlAsH III ΔALL variant (Fig. 4a). Conformational dynamics of DVL3 ΔALL was lost but, interestingly, the FRET efficiency for all three conditions was low—suggesting that DVL3 ΔALL remains in the open rather than the closed conformation. To further analyze this phenomenon, we produced CK1ε-deficient (CK1ε$^{-/-}$) HEK293 cells using the CRISPR-Cas9 system (Fig. 4b). These cells (Fig. 4b) failed to respond to Wnt ligands as demonstrated by the lack of phosphorylation of DVL2 and DVL3, and pS1490-LRP6. DVL3 overexpression in CK1ε$^{-/-}$ cells failed to induce Wnt/β-catenin pathway activation monitored by TopFlash in the absence of exogenous CK1ε (Supplementary Fig. 4f). Importantly, the FRET efficiency of the DVL3 FlAsH III sensor in CK1ε$^{-/-}$ cells was low and CK1ε inhibition was unable to increase it as it did in HEK293 wt cells (Fig. 4c). These data suggest that DVL3 in the absence of CK1ε remains in an open (and non-phosphorylated) rather than a closed (and non-phosphorylated) conformation that is observed when CK1ε is present but inhibited by the CK1δ/ε inhibitor PF670462. One explanation can be non-specific effects of CK1δ/ε inhibitor PF670462, unrelated to CK1ε inhibition. To exclude this possibility, we overexpressed *d*ominant *n*egative (dn) CK1ε mutant P3, which efficiently binds to DVL3 (ref. [33]). The presence of dnCK1ε as well as wild-type CK1ε in combination with PF670462 inhibitor resulted in the significantly increased FRET signal of the DVL3 FlAsH III sensor (Fig. 4d). This confirmed that in presence of inactive CK1ε, DVL3 adopts a more compact conformation which is not the case when CK1ε is absent from the DVL3 complex. In summary with other results, it suggests that CK1ε has a dual role: (i) it retains DVL3 in a compact conformation by physical interaction and (ii) triggers a phosphorylation-induced open conformation of DVL3, when activated.

C-terminus-PDZ interaction controls conformations of DVL3. The recent study by Lee et al.[16] has proposed that DVL can switch between open and closed conformations. Using mouse Dvl1 they demonstrated that the closed conformation results from the binding of the conserved terminal 7 aa of the C-terminus (EFFVDIM sequence) to the PDZ domain (for schematics see Fig. 5a). Therefore, we asked whether this interaction contributes at least partially to the DVL3 conformational dynamics detected by our FRET sensor. Indeed, DVL3 lacking the very C-terminal part (the ΔC variant of DVL3 FlAsH III, which lacks aa 699–716; Supplementary Fig. 5b) showed significantly lower basal FRET efficiency than wt DVL3 FlAsH III (Fig. 5b). This suggests that the interaction between the C-terminus and PDZ may play a role in the conformational sampling of DVL.

Our next question was whether CK1-mediated phosphorylation in the PDZ domain and the C-terminus contributes to the conformational dynamics induced by CK1ε. First, we employed the MS/MS approach to analyze the phosphorylation of DVL3 induced by CK1ε in the PDZ domain of DVL3. Two phosphorylation sites highly induced by CK1ε were found in the PDZ domain: S268 and S311 in DVL3 immunoprecipitated from HEK293 cells (Fig. 5c) as well as in the in vitro purified full-length DVL3 phosphorylated by recombinant CK1ε in vitro (Fig. 5d). Furthermore, C-terminal S700 in DVL3 was found to be constitutively phosphorylated in HEK293 cells, as previously mentioned (Supplementary Fig. 4b). All three serine residues are fully conserved (Supplementary Fig. 5a, b). To investigate how the phosphorylation status of S268 and S311 in the PDZ domain influences DVL3 conformation, we created ECFP-DVL3 FlAsH III wt construct that was either non-phosphorylatable (S268A/S311A) or phosphorylation-mimicking (S268E/S311E). DVL3 FlAsH III S268A/S311A variant showed higher FRET efficiency indicating the closed conformation, whilst the FRET efficiency of the DVL3 FlAsH III S268E/S311E variant was lower (Fig. 5e), suggesting predominantly the open conformation for this variant. Intriguingly, the FRET efficiency was not significantly affected by the CK1ε inhibition or CK1ε overexpression in any of these variants (Fig. 5e).

Altogether, these data suggest that the CK1-induced phosphorylation of S268 and S311 can control the conformation of DVL3. This led us to investigate whether this regulation can be explained at the atomic level as a result of weaker binding of the (phosphorylated) C-terminus to the PDZ domain phosphorylated at S268 and/or S311.

PDZ domain phosphorylation regulates the C-terminus binding. We performed in silico simulations to investigate the interaction between the C-terminal peptide of human DVL3 (aa 709–716) and human DVL3 PDZ (aa 245–338) domain. We found that the C-terminal peptide binds to the PDZ domain in



the canonical binding groove (primary binding site[34]) forming an antiparallel β-strand with β2-strand of the PDZ domain (Fig. 6a). Moreover, this interaction extends to the β2–β3 loop, where S268 via its hydroxyl forms a hydrogen bond with the carboxyl group of E710 (Fig. 6b). β2–β3 Loop has been reported as an additional binding site in the PDZ domains of several other proteins[34–36] but in DVL, such interaction (S268-E710) has not yet been reported.

To estimate the effect of phosphorylation on peptide binding in silico, we investigated the interaction between the peptide and the

Fig. 6 PDZ phosphorylation regulates the interaction with the DVL C-terminal peptide. **a** Structure of DVL3 C-terminal peptide (aa 702–716; green) bound to the PDZ-binding pocket (PDZ domain from DVL3, aa 245–338; in silico α-helices in dark and β-strands in light gray) observed in the in silico simulations. S268, S311, and C-terminus residue E710 participating in hydrogen bonds (dotted line) are highlighted with a stick model. **b** The matrix of mean interaction energy between each residue of PDZ domain aa 245–338 (*x*-axis) and DVL C-terminal peptide aa 709–716 (*y*-axis). Strength of the attraction of PDZ wt and C-terminus is depicted in white-black gradient. **c**–**e** The differences in mean interaction energies shown as a difference from wild-type PDZ wt: **c** PDZ (S268E/S311E), **d** PDZ (phospho-S268/phospho-S311), and **e** PDZ (S268A/S311A) between PDZ mutants. The change in the interactions are depicted in green (stronger) or red (weaker). Interactions of protein/peptide end caps are also displayed in the matrix. **f**–**i** NMR titrations of the DVL2 PDZ wild type and PDZ phosphomimicking variant (S286E/S329E; corresponding to S268 and S311 in DVL3) with DVL C-terminal peptide. **f** Overlay of ^{1}H, ^{15}N HSQC spectra for each titration point. Arrows indicate selected residues that exhibit fast exchange properties on the NMR timescale and gray boxes selected residues that exhibit intermediate-to-fast exchange properties on the NMR timescale. **g** Mapping of chemical shift perturbations on DVL-peptide (in green) structure from PDB database (PDB ID: 3CCO). Fast exchange residues colored using a gradient from white to red according to chemical shift perturbation and intermediate exchange residues that go beyond detection at the end of the titration colored in gray. S286, or E286 substitution, of DVL2 are highlighted by black arrow. **h** Binding isotherms for three residues that experience fast exchange during NMR titration. The apparent K_D values represent the mean with the standard deviation for the three cross-peaks analyzed. **i** Experimental line shapes during the titration for two selected residues that experience intermediate-to-fast exchange. **j** SPR sensograms of DVL C-terminal peptide binding to wild type and S286E/S329E PDZ from DVL2 and the corresponding binding isotherms fitted to a one-site binding model. RU stands for response units, AU for arbitrary units

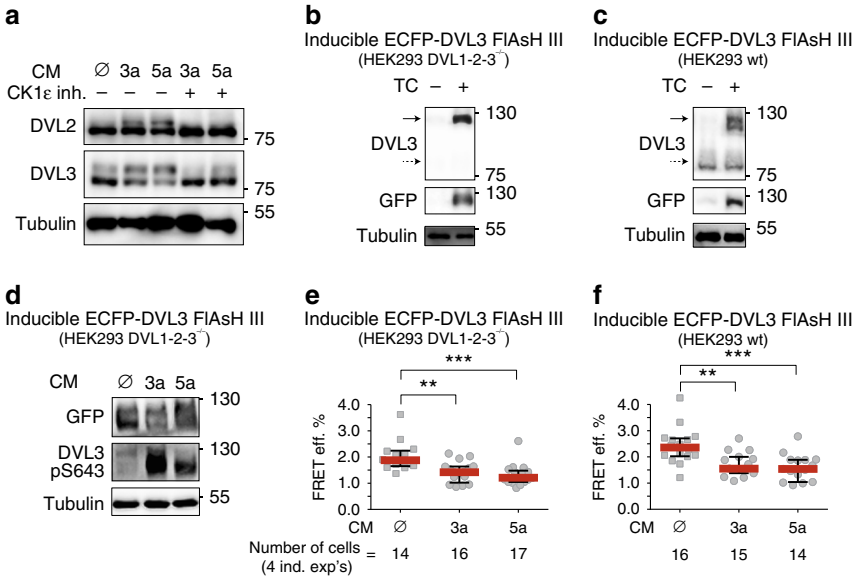
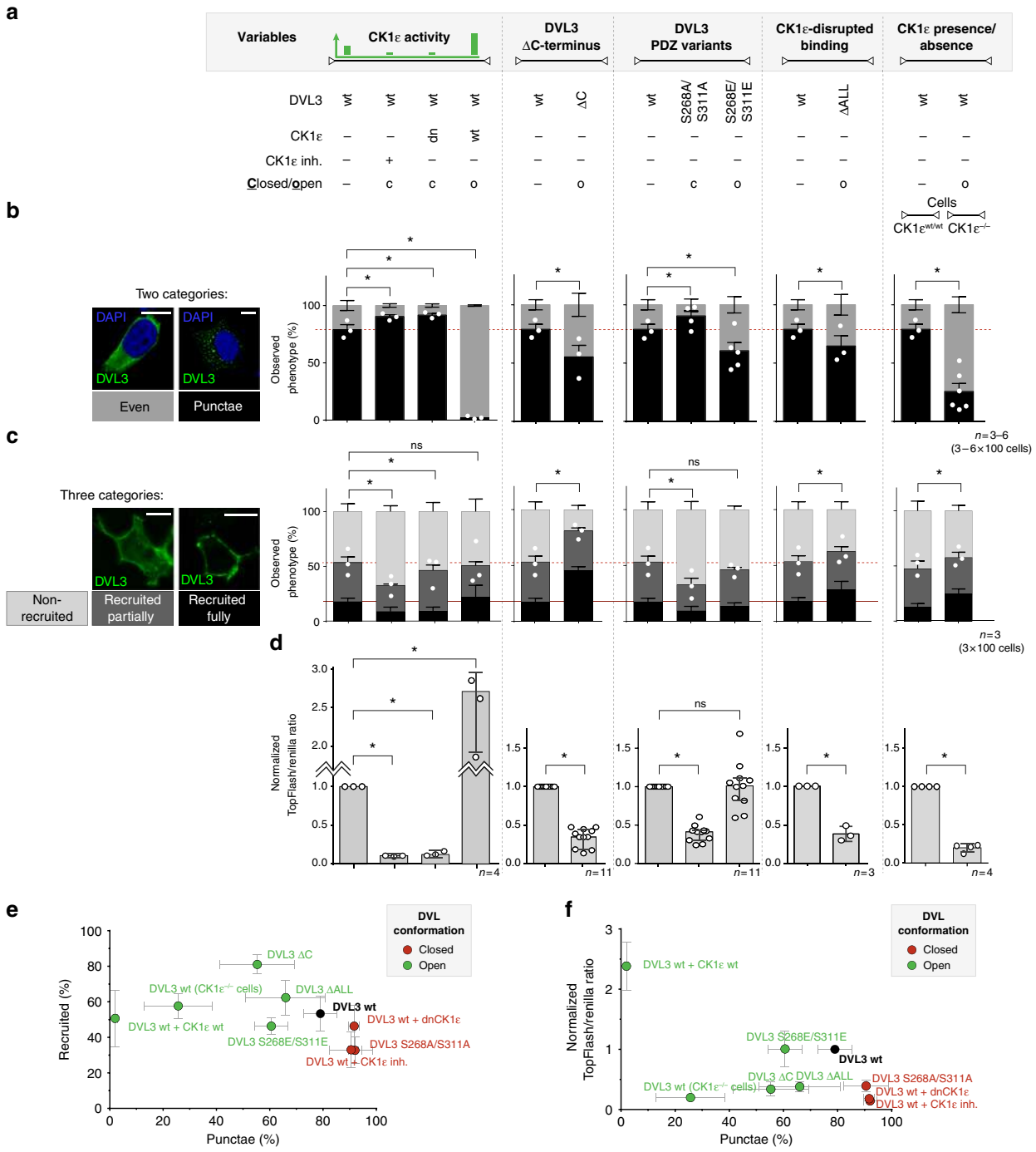


Fig. 7 Wnt ligands promote open conformation of DVL3. **a** Western blot analysis of the effect of the Wnt-3a (3a) and Wnt-5a (5a) ligands on the phosphorylation of endogenous DVL2 and DVL3 in HEK293 wt cells. **b**, **c** Western blot analyses of samples from two stable cell lines (derived from HEK293 wt and HEK293 DVL1/2/3 triple knockout[22] cells) that inducibly under tetracycline-controlled promoter express ECFP-DVL3 FlAsH III sensor. Dashed arrow indicates endogenous DVL3; full arrow indicates ECFP-DVL3. **d** Western blot analysis of the effect of the Wnt-3a (3a) and Wnt-5a (5a) ligands on the phosphorylation status of ECFP-DVL3 FlAsH III in HEK293 DVL1/2/3 triple knockout[22] cells. **e**, **f** Measurements of the intramolecular FRET efficiency of the endogenously expressed DVL3 FlAsH sensors III after the treatment with Wnt ligands in HEK293 wt and HEK293 DVL1/2/3 triple knockout[22] cells. One data point corresponds to one analyzed cell; datapoints from four independent transfections were merged. Data in **e** and **f** represent median ± interquartile range. Statistical significance in **e** and **f** was analyzed by one-way ANOVA test with Gaussian distribution and Tukey's post-test (*, $p \le 0.05$; **, $p \le 0.01$; ***, $p \le 0.001$, ****, $p \le 0.0001$; ns, not significant, $p > 0.05$). CM conditioned medium, Inh. inhibitor, TC tetracycline

DVL3 PDZ with two phosphorylated serines (pS268/pS311), phosphomimic S268E/S311E, and non-phosphorylatable S268A/S311A. In all cases, the peptide remained bound to the PDZ during the entire simulation (500 ns). However, the enthalpic interaction between the PDZ and the peptide were weaker for both phosphorylated and the mimetic variant by roughly 20% when compared to PDZ wt (Fig. 6c, d). In contrast, the alanine double mutant showed similar interaction strength as wt (Fig. 6e). The changes to the interaction strength were mostly situated in proximity to mutated/phosphorylated sites in the β2–β3 loop with a minor change in the primary carboxyl binding site (see Fig. 6b). The observed interaction strengths correlate with the preference of each PDZ mutant for open or closed conformation observed in the in vivo experiments and provide a candidate mechanism supporting FRET experimental observations.

In order to address the possible role of DVL3 phosphorylation-driven regulation involving S268 and S311 in vitro, we performed NMR structural characterization (Fig. 6f–i). For this purpose, we exploited the well-characterized PDZ domain of human DVL2 (aa 265–361)[14], where S286 and S329 correspond to S268 and S311 of human DVL3. Investigation of chemical shift differences revealed that the phosphorylation-mimicking substitutions (S286E/S329E) do not affect the PDZ fold[16] (Supplementary Fig. 6a–c). Next, we performed NMR titrations with a phosphorylated (pS700) C-terminal peptide ligand derived from DVL3 (aa 698–716). The phosphorylation variant was chosen as the most physiological, since S700 appeared as constitutively phosphorylated (see Supplementary Fig. 4b). The peak trajectories induced by peptide binding are linear and follow the same course for both PDZ wt and phosphorylation-mimicking variant



indicating a similar binding mode. Of the two serine residues, only S286 is affected by peptide binding and clearly experiences larger perturbation than its glutamic acid counterpart (Fig. 6f, Supplementary Fig. 6d). This is in line with in silico data, proposing that S286 in DVL2 (corresponding to S268 in DVL3) is directly involved in contacts with the C-terminal peptide.

Intermediate and intermediate-to-fast exchange peaks correspond to PDZ residues of the binding groove (Fig. 6g). These residues experience severe line broadening throughout the course of the ligand titration in PDZ wt when compared to S286E/S329E variant (Fig. 6i). Fast exchange peaks correspond to residues at the groove periphery. Their quantitative analysis shows that the phosphorylation-mimicking variant has a two-fold weaker affinity for the peptide (Fig. 6h). In principle, the chemical shift perturbations are larger for PDZ wt, indicating that the equilibrium is shifted further to the bound state when compared to the PDZ phosphorylation-mimicking variant (Supplementary Fig. 6d). As an independent confirmation of this conclusion, we also performed surface plasmon resonance (SPR) binding experiments (Fig. 6j). The apparent dissociation constants derived from SPR are in line with the NMR observations and reported a two-fold lower peptide affinity for the PDZ phosphorylation-mimicking variant.

In summary, the biophysical data and the in silico simulations indicate a phosphorylation-driven mechanism that modulates the binding of DVL C-terminus to PDZ domain. In combination

Fig. 8 DVL conformations correlate with DVL subcellular localization and membrane recruitment. **a** Conditions and variants with the defined conformational state identified by FlAsH-based FRET subjected to analyses in **b**–**d**; o indicates open, c indicates closed. **b** ECFP-tagged DVL3 was transfected to HEK293 wt cells as indicated. Subcellular localization was analyzed by anti-GFP immunostaining. Based on the pattern of ECFP-DVL3, cells were classified to have either even or punctate localization (plotted as white dots) of DVL3. Typical examples are shown in the left. Data represent average from three independent transfections with 100 cells counted in each condition per transfection. Identical control condition is shown several times for better clarity. **c** FLAG-tagged DVL3 and FZD6-mCherry were transfected in HEK293 DVL1/2/3$^{-/-}$, wt, and CK1ε$^{-/-}$ cells as indicated. Subcellular localization was analyzed by anti-FLAG immunostaining. Based on the membrane localization of FLAG-DVL3, cells were classified to possess DVL3 either non-recruited, partially recruited, or fully recruited (the sum of partial and full DVL3 recruitment plotted as white dots). Typical examples are shown in the left. Data represent average from three independent transfections with 100 FZD6-mCherry-positive cells counted in each condition per transfection. Identical control condition is shown several times for better clarity. **d** Analysis of the Wnt/β-catenin downstream signaling monitored by Dual Luciferase TopFlash/Renilla Reporter Assay in HEK293 cells. Data in **b**–**d** represent mean ± S.D. and statistical significance in **d** was analyzed by one-way ANOVA test with Gaussian distribution and Tukey's post-test. For more details about the statistics used in Fig. 8b and c, please see the appropriate section in Methods (*, $p \leq 0.05$; ns, not significant, $p > 0.05$). **e**, **f** Data obtained in **b**–**d** were plotted in the 2D graphs. Red color indicates the closed conformation; green indicates open. Error bars show S.D. for each parameter. % of recruited is a sum of partially recruited and fully recruited in **c**. DVL variants that are closed form a distinct population when membrane recruitment and subcellular localization is considered (**e**) but not when the TopFlash assay is included (**f**, and Supplementary Fig. 7a); dn dominant negative

with other results in this study, we propose that such mechanism contributes to the CK1-dependent control of DVL conformational changes.

Wnt ligands promote open conformation of DVL3. It has been shown previously that stimulation with Wnt ligands activates CK1δ/ε that in turn phosphorylate DVL. Wnts from both major groups—i.e. capable to activate Wnt/β-catenin (i.e. Wnt-3a) or alternative non-canonical Wnt pathway (i.e. Wnt-5a)—trigger phosphorylation of DVL that is dependent on CK1δ/ε activity (Fig. 7a). Thus, we decided to test if/how Wnts change the conformation of DVL3. To do this, we generated two stable cell lines (derived from HEK293 wt and HEK293 DVL1/2/3 triple knockout[22] cells) utilizing tetracycline-controlled promoter driving expression of ECFP-DVL3 FlAsH III sensor (Fig. 7b, c). Under these conditions, the ECFP-DVL3 FlAsH III gets expressed at levels comparable to endogenous DVL3 (Fig. 7c), and in response to Wnt stimulation gets phosphorylated—this can be monitored either as a phosphorylation-dependent shift or by anti-phospho-S643 antibody (Fig. 7d). Analysis of FRET efficiency in both cell lines showed that basal FRET efficiency (≈2%) was comparable with the control conditions analyzed upon transient overexpression of ECFP-DVL3 FlAsH III (see Figs. 2 and 4). Importantly, treatment with both Wnt-3a and Wnt-5a resulted in significantly lower FRET, regardless of the cell line used (Fig. 7e, f). Altogether, these results show that stimulation with Wnt ligands changes DVL3 conformation similarly to the overexpression of CK1ε and suggest that the phosphorylation-driven conformational switch participates in the mechanism of Wnt-induced DVL activation.

Open DVL3 is evenly localized and better membrane recruited. Our analyses of DVL3 conformational dynamics by FRET identified multiple conditions that promote either open or closed conformations of DVL3 (see Figs 2, 4 and 5). These conditions/mutations (schematized in Fig. 8a) include changes in the CK1ε activity, presence/absence of DVL3 C-terminus, phosphomimicking/blocking mutations in the PDZ domain, capacity of DVL3 to interact with CK1ε, and presence or absence of CK1ε in cells. Altogether, we could identify five cases associated with open conformation and three cases associated with closed conformation. In order to link DVL3 conformational changes with its biological properties, we decided to perform a systematic correlative analysis. The following conditions were included as closed: DVL3 wt + CK1δ/ε inhibitor PF-670642, DVL3 wt + dnCK1ε, and DVL3 (S268A/S311A), or open: DVL3 wt + CK1ε wt, DVL3 ΔALL, DVL3 (S268E/S311E), DVL3 ΔC, and DVL3 wt in CK1ε$^{-/-}$ cells.

First, we analyzed the preferred subcellular localization of DVL3 (as introduced in Fig. 2c)[10,14]. As shown in Fig. 8b, the open variants of DVL3 are significantly more evenly distributed whereas closed DVL3 is more punctate. Next, we investigated if DVL3 conformation correlates with the capacity of DVL3 to be recruited by FZD6, a key Wnt receptor, to the plasma membrane. For this purpose, we co-expressed FZD6 and analyzed membrane recruitment of DVL3 in FZD6-positive cells by immunofluorescence[37]. Individual cells were categorized as: not recruited (Fig. 8b, no DVL3 in the membrane), partially recruited (clear membrane as well as cytoplasmic localization), and fully recruited (predominantly plasma membrane localization, Fig. 8c, left). This analysis showed that closed DVL3 is less efficiently recruited to the membrane by FZD6, whereas open DVL3 behaves either as a control or is recruited more efficiently (Fig. 8c). In the last readout we analyzed the capacity of DVL3 to activate downstream Wnt/β-catenin pathway monitored by TopFlash reporter assay (Fig. 8d). Under all conditions where DVL3 is closed—DVL3 (S268A/S311A), DVL3 wt + CK1δ/ε inhibitor, and DVL3 wt + dnCK1ε—no (or negligible) TopFlash/Renilla activation was observed. On the other hand, under conditions that favor the open conformation both activation as well as inhibition of the Wnt/β-catenin pathway was evident.

Next, we plotted the biological properties of individual open and closed cases in a single graph (Fig. 8e, f; Supplementary Fig. 7a). When both subcellular localization and FZD6-induced membrane recruitment were considered, open and closed conditions separated from wt DVL3 very clearly (Fig. 8e). Ability to activate TopFlash was observed only when CK1ε was co-expressed with wt DVL3. Thus, it seems that CK1ε activity represents an independent additional factor that controls the behavior of open DVL3 forms; i.e., open conformation is not sufficient (although it can be required) for DVL3 to activate downstream Wnt/β-catenin signaling. Altogether these data demonstrate a strong correlation between DVL3 conformations and their ability to be recruited to the receptor complex (FZD6) or localize evenly in a cell.

In the cell culture model we were unable to dissect the effects on the non-canonical Wnt/PCP pathway, and as such we turned into the *Xenopus* embryo model. Alterations in the Wnt/PCP pathway activity result in the convergent extension (CE) defects (Supplementary Fig. 7b, right). In order to avoid any artifacts, we tested the constitutively open and closed variants of xDvl3 based on point mutations or small deletions—namely open xDvl3 ΔC and xDvl3 (S267E/S310E) and closedxDvl3 (S267A/S310A).

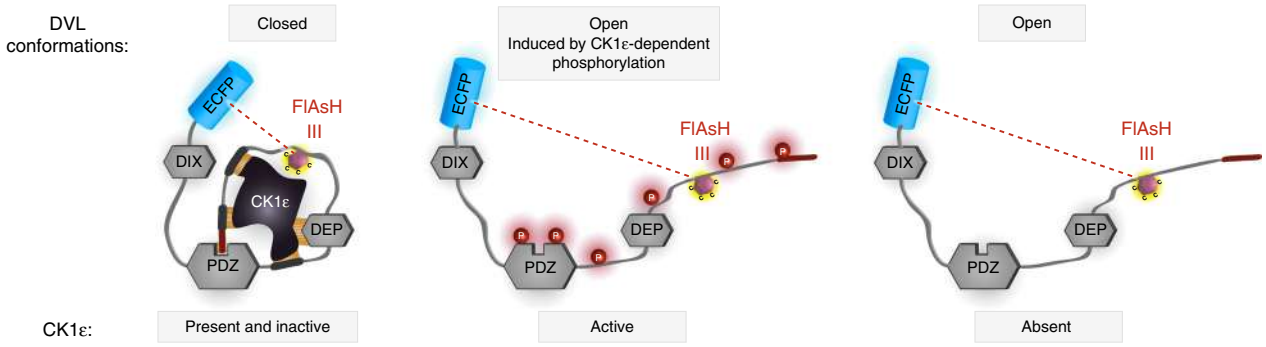


Fig. 9 Summary showing CK1ε role in DVL3 conformational dynamics. A summarizing model which proposes at least three DVL conformations in vivo: (i) a closed (CK1ε present and inactive), (ii) open (CK1ε active), and (iii) non-physiological open, which occurs when CK1ε is absent or the DVL-CK1ε interaction is disrupted. Position of insertion of FlAsH III binding tag is indicated. The CK1-induced phosphorylation events are depicted as P in red circle and the C-terminus of DVL as red thick line. The molecular distance analyzed in the FRET FlAsH sensor III is shown as a dashed red line; ECFP, enhanced cyan fluorescent protein

Phosphorylation sites in the PDZ domain are fully conserved between human and Xenopus (for alignment see Supplementary Fig. 5a) and hDVL3 S268/S311 corresponds to xDvl3 S267/S310. As shown in Supplementary Fig. 7b, open variants of xDvl3 showed similar potency to induce CE defects as wt xDvl3 with approximately 70–80% embryos displaying CE phenotype, whereas the closed xDvl3 (S267A/S310A) showed significantly reduced capacity to induce CE/NT (neural tube) defects with less than 50% affected embryos. These results correlate strongly with the capacity of the corresponding human DVL3 variants to be recruited to the membrane by FZD6 (Fig. 8c).

Discussion

In this study we establish an FlAsH-based single-cell FRET tool for the investigation of DVL conformations in intact living cells. Using this approach, we identify a key role of CK1ε, a major DVL kinase, in controlling DVL conformation in vivo.

The concept of DVL conformational sampling was first introduced by Jung et al.[38]. They proposed that the C-terminus of DVL is a ubiquitin-binding domain that can recognize K63-linked ubiquitin chains in the DEP domain to yield an inactive, closed, conformation[38]. Subsequent studies of Zheng group[16,17], in vitro proposed an alternative open/closed conformational switch mediated by docking of the C-terminus to the PDZ domain of DVL. The FlAsH-based FRET DVL3 sensor technique allows us to identify a key role of phosphorylation in the regulation of this process. Specifically, we identify two CK1ε-induced phosphoserine residues (S268 and S311 in human DVL3) in the PDZ domain whose mutation dramatically interferes with the DVL3 conformational sampling, and further supports the C-terminus–PDZ interaction model[16,17].

Previously, we have shown[28] that efficient interaction of DVL3 and CK1ε requires multiple DVL3 regions. The DIX domain is the only region of DVL3 which seems to be dispensable for high affinity interaction[28]. The DEP domain appears to be the most important structured domain involved in CK1ε binding[28,39]; here we identify three regions in the IDRs of DVL3 that are involved in CK1ε interactions. DVL3 lacking these three short sequences is not only inefficient to bind CK1ε but it is also not amenable to conformational rearrangements. Interestingly, such DVL3 variant is found counterintuitively in the open conformation. The same applies when CK1ε is eliminated in the CK1ε$^{-/-}$ cell line. This suggests that CK1ε can also have an important scaffolding function required to retain DVL3 in the closed conformation. This view is schematized in Fig. 9. Similar function has been observed for other kinases that regulate biological processes not only by their catalytic activity but also by their scaffolding function[40]. In principle, the requirement for CK1 as a scaffold can help to elucidate earlier findings shown by genetic experiments in *Drosophila* that Wnt/PCP pathway requires CK1ε but not its kinase activity (in contrast to Wnt/β-catenin pathway that requires both)[41].

The key question is how DVL conformation affects the formation of Wnt signaling complexes. We can show that both Wnt-3a, which triggers Wnt/β-catenin pathway, and Wnt-5a ligand, which activates non-canonical Wnt pathway, promote open conformation of DVL3. Both these ligands activate CK1ε, which results in a very similar or identical phosphorylation of DVL[24]. At the same time, we could correlate open conformations of DVL3 with the efficient FZD6-dependent membrane recruitment of DVL3. A similar correlation was reported recently using XDsh, XFz7, and zebrafish embryonal cells[17]. Based on these facts, we propose a model where Wnt-induced activation of CK1ε leads to the phosphorylation of S268/S311 in the PDZ domain and an open DVL conformation; i.e. open DVL3 then gets more efficiently bound to FZD. The molecular mechanism behind the more efficient recruitment to FZD-centered receptor complex is unknown but several candidate mechanisms reported in the literature can be considered for further testing. First, it is possible that in a closed conformation (i.e. when the binding groove of PDZ is occupied by DVL C-terminus), a conserved KTXXXW motif in the C-termini of FZDs that is capable of direct interaction with the PDZ[42,43] domain cannot interact with DVL. Second, the liberation of the DVL C-terminus in the open conformation provides a binding interface that can be recognized by proteins that facilitate receptor complex assembly. Such a candidate In the non-canonical Wnt pathway could be Ror2, which binds the C-terminus of DVL and more efficiently recognizes DVL phosphorylated by CK1 (ref. [44]). Third, our data show that the closed variants of DVL3 are more punctate whereas open variants are more evenly distributed. These phenotypes approximate the multimerization properties of DVL mediated by DIX and DEP domains[10,14]. In all our conditions, DVL3 has intact DIX and DEP domains, which suggests that the conformation itself affects multimerization and/or biophysical properties of DVL. One can speculate that closed DVL3 represents an inactive and immobile pool of DVL whereas open DVL is more mobile. Existing reservoirs of inactive or autoinhibited DVL have been proposed before[17,45]—for example: DVL-binding protein nucleoredoxin (NXN) can recognize such a pool and prevent it from degradation[45,46]. The key function of CK1 upon activation of Wnt pathway can then be mobilization of DVL from the inactive pool to be followed by FZD-mediated events—this distinction is

in line with earlier findings showing that the action of CK1ε and FZD on DVL is molecularly distinct[29].

Taken together, we provide evidence that Wnts and CK1ε regulate the conformation of DVL3, and that DVL3 in the open conformation, induced by these stimuli, is more mobile and more efficiently recruited to FZD. The FlAsH FRET DVL3 sensor technique provides the tool for the follow-up studies that can shed light on the diverse DVL functions at the membrane, in the centrosome, in the nucleus, and mainly on the regulation of DVL function in the Wnt/β-catenin and Wnt/PCP signaling pathways.

Methods

Cell culture and transfection and mutagenesis and treatments. HEK293 wild type (ATCC-CRL-11268), DVL1-2-3$^{-/-}$ (published in ref. [22]), ECFP-DVL3 FIII and CK1ε$^{-/-}$ cell lines (both derived from the aforementioned HEK293 wild-type cells) were grown at 37 °C and 5% (vol/vol) CO_2 in Dulbecco's modified Eagle's medium (DMEM; Gibco #41966-029), 10% (vol/vol) fetal bovine serum (Gibco, #10270), 2 mM L-glutamine (Life Technologies, #25030024), and 1% (vol/vol) antibiotics (penicillin/streptomycin; Hyclone-Biotech, #SV30010).

For immunofluorescence, cells were seeded (40,000 per well) on 24-well plates with gelatin-coated coverslips and transfected, as indicated, with 0.1 µg of each corresponding plasmid (0.3 µg of mCherry-FZD6) for 6 h. For western blotting and Dual Luciferase TopFlash/Renilla Reporter Assay, cells were seeded (60,000 per well) directly on 24-well plates with 0.1 µg of each corresponding plasmid (if not indicated otherwise) for 6 h. For coimmunoprecipitation, cells were seeded (4.4×10^6) on 10-cm dishes and transfected with 1 µg of each corresponding plasmid. The next day after seeding, cells were transfected using polyethyleneimine (PEI) in a stoichiometry of 6 µl of PEI per 1 µg of DNA for 6 h and were harvested 24 h after transfection. The following plasmids used in HEK293 cells have been described previously: FLAG-DVL3 wt[47], CK1ε wt[48], mCherry-FZD6 (ref. [37]), and dominant-negative CK1ε P3 (ref. [33]). The ECFP-DVL3 wt plasmid was made by Gateway Technology (via LR recombination according to the manufacturer's instructions; Thermo Fisher Scientific, #11791020) from the donor plasmid pDONR221 DVL3 (DNASU, #FLH178665.01x) and the destination plasmid pdECFP (LMBP, #4548). The HA-DVL3 wt plasmid was made by Gateway Technology from the donor plasmid pDONR221 DVL3 (DNASU, #FLH178665.01x) and the destination plasmid pcDNA3.1 N-term HA (gift from Erich Nigg).

Mutagenesis reactions were performed using the QuikChange II XL Site-Directed Mutagenesis Kit following the manufacturer's instructions (Agilent Technologies, #200521). All mutations described in this study were verified by Sanger sequencing. All primers used for the site-directed mutagenesis are listed in Supplementary Table 1. All plasmids and more detailed information are available upon request.

Cells were treated as indicated by 10 µM CK1δ/ε inhibitor PF-670642 (Santa Cruz, #sc-204180A) overnight; 1 µM LGK974 Porcupine inhibitor (Stem RD, #974-02) overnight; 1 µg/ml of tetracycline for 4 h; and by Conditional Media (CM) for 3 h (WB, FlAsH-based FRET approach) or 8 h (TopFlash/Renilla Dual luciferase assay), which were obtained from the control or Wnt-3a/Wnt-5a stably overexpressing L-cells mouse cell line (L-Wnt3a; ATCC-CRL-2647 and L426 Wnt5a cells; ATCC-CRL-2814) according to the ATCC instructions.

CRISPR-Cas9 generation of HEK T-REx-293 CK1ε$^{-/-}$ cell line. The sequence of gRNA targeting CSNK1E gene was chosen according to Broad Institute CRISPRko designing tool as follows: gRNA (CSNK1E)—AAGTTCTACAAGATGATGCA (GGG) (SfaNI restriction site indicated, PAM sequence in brackets).

The following plasmids were used for gRNA cloning: pSpCas9(BB)-2A-GFP (PX458) (Cas9 from *S. pyogenes* with linked EGFP and cloning backbone for gRNA, Addgene plasmid #48138)[49] and pU6-(BbsI)_CBh-Cas9-T2A-mCherry (Cas9 from *S. pyogenes* with linked mCherry and cloning backbone for gRNA, Addgene plasmid #64324)[50].

T-REx-293 cells (Invitrogen) were cultured according to the manufacturer's instructions and transfected by Lipofectamine® 2000 DNA Transfection Reagent (Thermo Fisher Scientific) with #48138 (cloned in gRNA for CSNK1E). Cells were subsequently subjected to FACS (fluorescence activated cell sorting) where simultaneously GFP (green fluorescent protein) and mCherry-positive cells were seeded in 96-well plates in one cell per well format.

Monoclonal cell lines were backed up and subsequently screened by western blot using CK1ε antibody (Santa Cruz, #sc-6471). For verification and specification of modification were above-mentioned PCR products cloned into pcDNA3 plasmid (Invitrogen) and sequenced using CMV forward primer: CGCAAATGGGCGGTAGGCGTG.

Generation of HEK T-REx-293 ECFP-DVL3 FlAsH III cell lines. cDNA encoding ECFP-DVL3 FlAsH III was amplified by Pfx50 DNA Polymerase (Thermo Fisher Scientific, #12355012) using primers containing sequences recognized by *Hin*dIII (forward primer TATAAGCTTatggtgagcaaggg; *Hin*dIII site underlined) and *Xba*I (reversed primer AATTCTAGAtcacatcacatccacaaa; *Xba*I site underlined) restriction enzymes (Thermo Fisher Scientific, #ER0501 and #ER0681), extracted from gel using GeneJET Gel Extraction Kit (Thermo Fisher Scientific, #K0691) and ligated into *Hin*dIII- and *Xba*I-linearized pcDNA4/TO (Thermo Fisher Scientific, #K102001) vector using T4 DNA Ligase (New England Biolabs, #M0202) according to the manufacturer's protocol. Then competent *E. coli* DH5α bacteria were transformed with ligation mix using standard heat-shock protocol, individual bacterial colonies were picked, cultured, and plasmid DNA was isolated by Gen-Elute Plasmid DNA Miniprep Kit (Sigma-Aldrich, #PLN70). Plasmids were checked by Sanger sequencing (forward sequencing primer ATTGACG-CAAATGGGCG; reverse GCCTTCCTTGACCCTGGA) for the presence of the ECFP-DVL3 FlAsH III wt cDNA. Next, T-REx-293 cells (Invitrogen) and their derivate HEK293 DVL1-2-3$^{-/-}$ cell line[22] were transfected with pcDNA-4-TO-ECFP-DVL3 plasmid using PEI. Twenty-four hour post-transfection culture medium was changed for the new one supplemented with 200 µg/ml for generation of stable cell lines. After selection clonal lines were obtained using the serial dilution method and analyzed by western blot using the GFP/DVL3 antibodies for the inducible transgene expression by treatment with 1 µg/ml of tetracycline.

Dual luciferase TopFlash/Renilla Reporter Assay. Cells were transfected 24 h after seeding with 0.1 µg of Super8X TopFlash construct, 0.1 µg of Renilla luciferase construct, and 0.1 µg of corresponding plasmids per well for 6 h in a 24-well plate. As for the ECFP-DVL3 FIII wt stable cell lines, transfection was followed by 4 h induction of tetracycline (1 µg/ml) and 8 h treatment of CM:DMEM (1:1 ratio). For performing this assay, Promega Dual Luciferase assay kit (Promega, #E1910) was used according to the manufacturer's instructions. Luminescence was measured by a Hidex Bioscan Plate Chameleon Luminometer. Results are depicted as the ratio of TopFlash and Renilla signal (TopFlash fold induction), which was normalized for each experiment to certain control column (normalized TopFlash fold induction). Data were analyzed by MS Excel 2007 and GraphPad Prism 6 and results were shown as the means ± S.D. (number of experiments indicated ad hoc).

Western blot and sample preparation. Coimmunoprecipitation protocol was used modified from Bryja et al.[29] DMEM was removed after 24 h, cells were washed by phosphate-buffered saline (PBS) and 1 ml of cold NP40-based lysis buffer (50 mM Tris buffer, pH 7.4; 300 mM NaCl; 1 mM EDTA; 0.5 % NP40) supplemented with 1× protease inhibitors (Roche Applied Science, #11836145001), 0.1 mM DTT and 1× phosphatase inhibitors (Calbiochem, #524625) was used per one 10-cm dish. The lysate was collected after 15 min of lysis on 4 °C and was cleared by centrifugation at 16.1 r.c.f. for 30 min. Samples with the antibody (1.25 µg of antibody per sample) were incubated on carousel overnight. In the morning, 30 µL of G protein-Sepharose beads (GE Healthcare, #17-0618-05)—previously blocked overnight in NP40-based lysis buffer supplemented with 5% bovine serum albumin (BSA) and in the morning equilibrated in BSA-free complete NP40-based lysis buffer—were then added to samples. After 4 h of incubation on carousel, samples were washed four times, 40 µL of 2× Laemmli buffer were added, and samples were boiled. The antibody used for immunoprecipitation was anti-CK1ε (BD Biosciences, #610445).

Blot and sample preparation were performed as previously described[51] and developed using chemiluminescence documentation system FusionSL (Vilber-Lourmat). The antibodies used were: anti-FLAG M2 (1:1000; Sigma-Aldrich, #F1804), anti-FLAG (1:1000; Sigma-Aldrich, #F7425), anti-DVL3 (1:1000; Santa Cruz, #sc-8027), anti-DVL3 (1:1000; Cell Signaling, #cs-3218S), anti-DVL2 (1:1000; Cell Signaling, #cs-3216S), anti-CK1ε (1:500; BD Biosciences, #610445), anti-CK1ε (1:500; Santa Cruz, #sc-6471), anti-GFP (1:2000; Fitzgerald, #20R-GR-011), anti-LRP6 pS1490 (1:250; Cell Signaling, #cs-2568), anti-tubulin (1:2000; Sigma-Aldrich, #T6199), anti-DVL3 phospho S643 (ref. [29]) (1:1000), and anti-DVL3 phospho S697 (ref. [52]) (1:1000). All relevant uncropped blots are available in the Source Data file.

Immunofluorescence. Cells were seeded on gelatin-coated coverslips in 24-well plates and were transfected the next day. Cells were fixed 24 h later in fresh 4% paraformaldehyde, permeabilized with 0.5% Triton X-100, blocked with PBS/BSA/Triton/Azide buffer (PBTA) [3% (wt/vol) BSA, 0.25% Triton, 0.01% NaN_3] for 1 h, and incubated overnight with primary antibodies (in dilution 1:500) in PBTA at 4 °C. The next day, the coverslips were washed in PBS and incubated with secondary antibodies conjugated to Alexa Fluor 488 (Invitrogen A11001) or/and Alexa Fluor 594 (Invitrogen A11058) (all in dilution 1:1000), washed with PBS [in case of samples containing FLAG-DVL3 WT and derived variants stained with DAPI (1:5000)], and all coverslips were mounted on microscopic slides. Cells were then visualized on an Olympus IX51 fluorescent microscope using ×40 air or ×100 oil objectives and/or an Olympus Fluoview 500 confocal laser scanning microscope IX71 using ×100 oil objective. One hundred positive cells per experiment ($n = 3$) were analyzed and scored according to their phenotype into two/three categories (punctae/even); or non-recruited/partially/fully recruited for FZD6-dependent recruitment of DVL3. The antibodies used were as follows: anti-FLAG M2 (Sigma-Aldrich, #F1804), anti-DVL3 (Santa Cruz, #sc-8027), anti-CK1ε (Santa Cruz, #sc-6471), anti-RFP (Chromotek, #5h9) and anti-GFP (Fitzgerald, #20R-GR-011).

***Xenopus laevis* embryos**. The work with *Xenopus laevis* was carried out according to the German animal use and care law (Das Tierschutzgesetz) and approved by the local authorities and committees (animal care and housing approval: I/39/EE006, Veterinäramt Erlangen; animal experiments approval: 54-2532.2-8/08, German state administration Bavaria/Regierung von Mittelfranken). The generation and cultivation of *Xenopus* embryos was done in accordance with general protocols and staged after the normal table of Niewkoop and Faber (1994)[53].

RNA for microinjection was synthesized from plasmids pCS2 myc-xDvl3 (ref. [54]), pCS2 myc-xDvl3 ΔALL (lacking sequence corresponding to aa 338–350, 609–619, and 693–705 in xDvl3), pCS2 xDvl3 (S267A/S310A), pCS2 xDvl3 (S267E/S310E), pCS2 xDvl3 ΔC, and pCS2 xCK1ε[55], respectively, using the mMessage mMachine Kit (Ambion). Embryos were injected at the four-cell stage either into the two ventral blastomeres for testing the constructs' ability to induce a secondary body axis. Injection amounts of synthetic RNA were 500 pg for myc-xDvl3 wild type and the myc-xDvl3 ΔALL variant, and 150 pg for xCK1ε RNA. For the analysis of morphogenetic defects, four-cell stage embryos were injected into the two dorsal blastomeres with 100 pg myc-xDvl3 RNA and the corresponding variant RNAs. After injection, the embryos were cultivated until they reached stage 26, fixed, and analyzed for axis duplication or morphogenetic defects, respectively.

FlAsH-based FRET. HEK293 cells were seeded onto round 24-mm coverslips, which were previously placed in six-well plates and coated with 200 µl of poly-D-lysine (1 mg/ml) for 20 min. Cells were transfected 16–18 h later using Effectene (Qiagen), according to the manufacturer's instructions. DMEM was replaced 6 h later and the analysis was done 24 h after transfection. As for the ECFP-DVL3 FIII stable cell lines: next day after seeding, protein expression was induced by 4 h treatment of tetracycline (1 µg/ml), which was followed by 3 h treatment of CM: DMEM (1:1 ratio).

FlAsH labeling of the DVL3 FlAsH sensors was performed as previously described[18,19]. In summary, transfected cells were washed once with Hank's Balanced Salt Solution (HBSS) containing 1.8 g/l glucose and then incubated at 37 °C for 1 h with HBSS supplemented with 500 nM FlAsH; 12.5 µM 1,2-ethanedithiol (EDT); and corresponding inhibitors. In order to reduce non-specific labeling, cells were then rinsed once with HBSS and incubated for 10 min with HBSS containing 250 µM EDT and corresponding inhibitors. Last, cells were washed twice with HBSS and maintained in DMEM medium with corresponding inhibitors (for the ECFP stable cells —CM:DMEM (1:1 ratio)).

Fluorescence imaging was performed as previously described[18,56]. To determine the FRET efficiency of the DVL3 FlAsH sensors, coverslips with the cells were mounted using an Attofluor holder (Molecular Probes) and placed on a Zeiss inverted microscope (Axiovert200), equipped with ×63 oil objective lens and a dual-emission photometric system (Till Photonics). Cells were excited at 436 ± 10 nm using a frequency of 10 Hz with 40 ms illumination time out of a total of 100 ms. Emission of ECFP (480 ± 20 nm) and FlAsH (535 ± 15 nm), and the FRET ratio (FlAsH/ECFP) were monitored simultaneously. Fluorescence signals were detected by photodiodes and digitalized using an analog-digital converter (Digidata 1440A; Axon Instruments)[18,56]. Fluorescence intensities data were acquired using Clampex software (Axon Instruments). During measurements, cells were maintained in imaging buffer; 5 mM of 2,3-dimercapto-1-propanol (also called British anti-Lewisite—BAL) was added to the cells approximately 40 s after the recording started. Recovery of ECFP fluorescence was monitored over time and FRET efficiency was calculated as described here[18,57]. One independent experiment represents approximately 4–6 repeats (i.e. single-cell FRET signals) for each condition.

Peptide array and other peptide reagents. The peptide library comprised of peptides (13-mer peptides, 10 residue-overlap) with sequences derived from the IDRs of DVL3. This peptide library comprised a set of non-phosphorylated peptides and a set of phosphorylated peptides, in which the phospho groups were located at those residues corresponding to the previously identified CK1ε's phosphorylation pattern[29] (Supplementary Fig. 3). The designed peptide library was generated by JPT Peptide Technologies GmbH (Berlin, Germany) in an array format, wherein the peptides were immobilized on a glass slide.

In each experiment, two glass slides were used: one was as blank and another was screened with the recombinant CK1ε. Both slides were blocked by treatment with SmartBlock solution (Candor Bioscience, #113125) for 1 h at 30 °C. Then, one slide was incubated overnight with 10 µg/ml of the recombinant human CK1ε (MyBioSource, #MBS964562) in SmartBlock solution at 4 °C and the second one with SmartBlock solution only. The next day, both slides were washed in TBS buffer (4 × 10 min, RT) and then incubated with 1 µg/ml of the primary anti-CK1 antibody in SmartBlock solution for 1 h at 30 °C. The slides were subsequently washed again in TBS buffer (4 × 10 min, RT), followed by incubation with 1 µg/ml of secondary fluorescent antibodies in SmartBlock solution for 45 min at 30 °C, and subsequent washing in TBS buffer (4 × 10 min, RT) and in water (4 × 10 min, RT). To remove the excess of water, the slides were spinned down using a swing rotor (1200 r.c.f. for 2 min, RT, slides placed in 50 ml falcons). Binding of CK1ε to the peptides was detected by reading the fluorescence intensity of the peptide spots. In these experiments, the primary anti-CK1ε antibodies used were from Santa Cruz Biotechnology (#sc-6471) and from BD Biosciences (#610445), and the secondary antibodies used were: Alexa Fluor 594 Goat Anti-Mouse IgG2a (γ2a) (Invitrogen, #A-21135) and Alexa Fluor 594 Donkey Anti-Goat IgG (H + L) (Invitrogen, #A-11058). The slide data were analyzed using the PepSlide Analyzer software (Sicasys).

In each experiment, the signal intensity from each peptide on the experimental set of two slides (i.e. the one used as blank and the one screened with CK1ε) was normalized to 0–100%, where the 100% value corresponded to the strongest signal detected for one of the peptide spots within the CK1ε-screened slide. The intensities measured for the peptide spots on the control slide—which was incubated with the detection antibodies only—were then subtracted from the intensities measured for the corresponding peptide spots on the CK1ε-screened slide. Each experiment was repeated 2–3 times (2–3 technical replicates per experiment) using the same anti-CK1ε antibody and average values for each antibody were added up and plotted in the graph. The signals with the highest intensity involving at least four consecutive peptides were considered as a putative CK1ε-binding site. The signals shown in the graph (Fig. 3b) correspond to the non-phosphorylated peptides and their numbering is arbitrary.

The peptides used in the fluorescent polarization (FA) binding assays were synthesized at Pepscan B.V. (Lelystad, Netherlands). The sequences of these peptides (indicated in Supplementary Fig. 4a) were amidated at their C-termini, and their N-termini were acylated with 6-aminocaproic acid (Ahx) modified with fluorescein isothiocyanate (FITC). The peptides used in the NMR/SPR studies were purchased from KareBay Biochem, Inc. (New Jersey, USA). The peptides for the NMR/SPR studies contained no terminal modifications.

Fluorescence anisotropy. Data were recorded at 37 °C using Tecan Infinite F500 (Tecan Systems, Germany). FITC-labeled peptides were dissolved in DMSO and added to 384-well plates (Corning) to a total volume of 20 µl (peptide final concentration was 50 nM). Incubation with recombinant His-tagged human CK1ε (MyBioSource, #MBS964562) was performed for 30 min at 37 °C. Then, fluorescence intensities were collected simultaneously by detection at 535/590 and 485/535 nm excitation/emission wavelengths, respectively, followed by adjustment to the blank controls.

Mass spectrometry. Samples were loaded onto SDS-PAGE gels, separated, and fixed with acetic acid in methanol, stained with Coomasie brilliant blue for 1 h and partially destained. Corresponding 1-D bands were excised. After destaining, the proteins in gel pieces were incubated with 10 mM DTT at 56 °C for 45 min. After removal of DTT excess samples were incubated with 55 mM IAA at room temperature in darkness for 30 min, then alkylation solution was removed, and gel pieces were hydrated for 45 min at 4 °C in digestion solution (5 ng/µl trypsin, sequencing grade, Promega, Fitchburg, Wisconsin, USA, in 25 mM AB). The trypsin digestion proceeded for 2 h at 37 °C on Thermomixer (750 r.p.m.; Eppendorf, Hamburg, Germany). Subsequently, the tryptic digests were subsequently cleaved by chymotrypsin (5 ng/µl, sequencing grade, Roche, Basel, Switzerland, in 25 mM AB) for 2 h at 37 °C. Digested peptides were extracted from gels using 50% ACN solution with 2.5% formic acid and concentrated in a speedVac concentrator (Eppendorf, Hamburg, Germany). The aliquot (1/10) of concentrated sample was transferred to LC-MS vial with already added polyethylene glycol (PEG; final concentration 0.001%)[58] and directly analyzed by LC-MS/MS for protein identification.

The rest of the sample was used for phosphopeptide analysis. Sample was diluted with acidified acetonitrile solution (80% ACN, 2% FA). Phosphopeptides were enriched using Pierce Magnetic Titanium Dioxide Phosphopeptide Enrichment Kit (Thermo Scientific, Waltham, Massachusetts, USA) according to the manufacturer's protocol and eluted into LC-MS vial with already added PEG (final concentration 0.001%). Eluates were concentrated under vacuum and then dissolved in water and 0.6 µl of 5% FA to get 12 µl of peptide solution before LC-MS/MS analysis.

LC-MS/MS analyses of peptide mixture were done using the RSLCnano system connected to an Orbitrap Elite hybrid spectrometer (Thermo Fisher Scientific) with ABIRD (Active Background Ion Reduction Device; ESI Source Solutions) and Digital PicoView 550 (New Objective) ion source (tip rinsing by 50% acetonitrile with 0.1% formic acid) installed. Prior to LC separation, tryptic digests were online concentrated and desalted using trapping column (100 µm × 30 mm) filled with 3.5-µm X-Bridge BEH 130 C18 sorbent (Waters). After washing of trapping column with 0.1% FA, the peptides were eluted (flow 300 nl/min) from the trapping column onto Acclaim Pepmap100 C18 column (3 µm particles, 75 µm × 500 mm; Thermo Fisher Scientific) by 65 min long gradient. Mobile phase A (0.1% FA in water) and mobile phase B (0.1% FA in 80% acetonitrile) were used in both cases. The gradient elution started at 1% of mobile phase B and increased from 1% to 56% during the first 50 min (30% in the 35th and 56% in 50th min), then increased linearly to 80% of mobile phase B in the next 5 min and remained at this state for the next 10 min. Equilibration of the trapping column and the column was done prior to sample injection to sample loop. The analytical column outlet was directly connected to the Digital PicoView 550 ion source.

MS data were acquired in a data-dependent strategy selecting up to top six precursors based on precursor abundance in the survey scan (350–2000 m/z). The resolution of the survey scan was 60,000 (400 m/z) with a target value of 1×10^6 ions, one microscan, and maximum injection time of 200 ms. High-resolution (resolution 15,000 at 400 m/z) HCD MS/MS spectra were acquired with a target

value of 50,000. Normalized collision energy was 32% for HCD spectra. The maximum injection time for MS/MS was 500 ms. Dynamic exclusion was enabled for 45 s after one MS/MS spectra acquisition and early expiration was disabled. The isolation window for MS/MS fragmentation was set to 2 *m/z*.

The analysis of the mass spectrometric RAW data files was carried out using the Proteome Discoverer software (Thermo Fisher Scientific; version 1.4) with in-house Mascot (Matrixscience; version 2.4.1) search engine utilization. MS/MS ion searches were done against in-house database containing expected protein of interest with additional sequences from cRAP database (downloaded from http://www.thegpm.org/crap/). Mass tolerance for peptides and MS/MS fragments were 7 ppm and 0.03 Da, respectively. Oxidation of methionine, deamidation (N, Q), and phosphorylation (S, T, Y) as optional modification, carbamidomethylation of C as fixed modification, and three enzyme miss cleavages were set for all searches. The phosphoRS feature was used for phosphorylation localization.

Quantitative information was assessed and manually validated in Skyline software (Skyline daily 3.1.1.8884).

Numerical data and statistics. As for numerical data, mean ± S.D. (or median ± interquartile range, respectively) are shown as depicted. No pairing one-way ANOVA test with Gaussian distribution (with Tukey's post-test) was used for comparison of more than two samples; for two-column statistics in Figs. 3e, g and 5b, unpaired two-tailed Student's *t*-test was used (*, $p \le 0.05$; **, $p \le 0.01$; ***, $p \le 0.001$, ****, $p \le 0.0001$; ns, not significant, $p > 0.05$).

Data on frequency of puncta vs. even (Fig. 8b) were analyzed using Generalized Linear Mixed Model (GLMM) with binomial distribution of dependent variable, treatment as fixed effect, and experiment as random effect variable, since data within the experiments are not independent. In case of evaluation of DVL-membrane recruitment results (Fig. 8c), the GLMM with multinomial distribution was used. Results were presented as odds ratios (OR) with 95% confidence intervals. Data were analyzed in R software. All *numerical* data were analyzed by MS Excel 2007 and arranged in the graphs using GraphPad Prism 6 software.

Multiple sequence alignment. The multiple sequence alignment of selected sequences was performed by ClustalW algorithm. Output alignment was refined manually using the BioEdit v7.0.1 sequence editor.

Protein expression and purification. The PDZ domain of human DVL2 wild type (aa 265–361) and phosphorylation-mimicking variant S286E + S329E were cloned into pET vector containing an N-terminal His_6-tag and a lipoyl domain separated by a Tobacco Etch Virus (TEV) protease digestion site from the N-terminus of the inserts. Both proteins were expressed at high yields in *E. coli* BL21 (DE3) strain (New England Biolabs, #C2527I). For the NMR studies, ^{13}C- and ^{15}N-labeled proteins were prepared by growing cells in minimal medium supplemented with $^{15}NH_4Cl$ (1 g/l) and $^{13}C_6$ glucose (2 g/l) as the sole nitrogen and carbon sources, respectively. Cells were grown at 37 °C to an OD_{600} of approximately 0.8 and protein expression was induced with 0.5 mM IPTG. Cells were lysed by sonication and centrifuged at 15,000*g*. Three purification steps were used for both proteins: Ni-NTA chromatography followed by overnight TEV cleavage to remove the N-terminal His_6-tag and the lipoyl domain, ion exchange chromatography, and size exclusion chromatography. Finally, the untagged PDZ domain wild type or phosphorylation-mimicking variant was concentrated and exchanged to buffer for NMR or binding experiments containing 20 mM HEPES (pH 6.8) and 50 mM KCl. Human CK1ε (aa 1–294) was also cloned into a pET vector and purified using three purification steps as described for the PDZ domain.

Full-length DVL3 used as a substrate in the kinase assay was produced as follows: Twin-Strep-tag N-terminally tagged DVL3 was expressed in HEK293 cells using transient transfection. Cells were harvested 48 h post-transfection, resuspended in a lysis buffer (50 mM Tris, 150 mM NaCl, 10 mM KCl, 10% glycerol, pH 8) with cocktail of protease inhibitors (#11836145001, Roche) and 0.2% NP40 (#74385, Sigma). The mixture was incubated for 20 min on ice and cell lysis was enhanced by sonication. Cell lysate was cleared by centrifugation at 100,000*g* for 45 min at 4 °C and supernatant was loaded on Strep-Tactin Superflow high capacity column (#2-1237-001, IBA) equilibrated in the lysis buffer. The column was washed in lysis buffer and the protein was eluted using lysis buffer supplemented with 3 mM desthiobiotin. Eluted proteins were concentrated to 1 mg/ml using protein concentrators (#88516; Thermo Fisher Scientific), flash frozen in liquid nitrogen, and aliquots were stored at −80 °C.

In vitro kinase assay. Full-length DVL3 protein and CK1ε kinase were dialyzed into phosphorylation buffer (50 mM sodium phosphate, pH 6.5, 50 mM KCl, 10 mM $MgCl_2$, 1 mM DTT, 1 mM EDTA). In each reaction, DVL3 and CK1ε were mixed in 2:1 ratio (containing ~0.5 µM DVL3, ~0.25 µM CK1ε kinase, and 1 mM ATP), and incubated at 25 °C for 16 h. Samples were subsequently analyzed by SDS-PAGE and mass spectrometry.

NMR spectroscopy. A 4D HC(CC-TOCSY(CO))NH and a 4D ^{13}C, ^{15}N edited HMQC-NOESY-HSQC experiments were recorded at CEITEC Josef Dadok National NMR Centre on a 700 MHz Bruker Avance III spectrometers equipped with $^1H/^{13}C/^{15}N$ TCI cryogenic probe head with *z*-axis gradients. 4D spectra were processed with SSA package[59] and analyzed with SPARKY. Chemical shift assignments were obtained automatically using 4D-CHAINS[60] and checked manually. NMR titrations were performed in series of ^1H, ^{15}N HSQC spectra using 100 µM of ^{15}N-labeled protein (wild type or phosphorylation-mimicking variant) and increasing amounts of C-terminal DVL peptide (pS700) consisted of aa 698–716 from DVL3 (stock concentration of 800 µM). Steady-state ^{15}N-^1H Nuclear Overhauser Effect values were measured under a steady-state condition with a 30 s interscan relaxation delay. Reference spectra and the spectra measured under steady-state conditions were measured in an interleaved manner.

Surface plasmon resonance. SPR experiments were performed on a Biacore T200 instrument (GE Healthare). The wild type and phosphorylation-mimicking variant of PDZ from DVL2 were immobilized on carboxy-methylated dextran matrix (CM5 chip S Series) using amino coupling. Non-reacted groups were blocked by 1 M ethanolamine. The blank channel was modified by NHS/EDC activation and subsequently blocked by ethanolamine. The C-terminal DVL peptide (aa 698–716 from DVL3; pS700) was injected at increasing concentrations and long enough to reach steady state. SPR responses were measured and corrected for the response of the blank. Data points were collected from each binding sensogram before the end of the injection (equilibrium) and the dissociation constants were estimated by fitting the response at equilibrium versus the C-terminal peptide concentration using the following equation: $R_{eq} = C^*R_{max}/(K_D + C)$, where R_{eq} is the binding level in Response Units at different C-terminal peptide concentrations (C), R_{max} the response at saturation (fitted), and K_D the apparent dissociation constant (fitted).

In silico simulation methodology. All simulations were done in Gromacs 5.1.2 simulation package[61]. Each simulation contained a PDZ domain in a salt solution of 150 mM NaCl. The initial structure of wild-type PDZ domain (aa 245–338) from DVL3 was constructed by homology modeling using MODELLER v9.11 (refs. [62,63]). As a template we used crystal structure of human DVL2 (PDB ID: 2REY) which share 96.28% sequence identity. Missing loops and one missing residue at C-terminus were added via MODELLER[64]. Quality of seven generated models was evaluated based on DOPE and GA341 score and the best structure has been selected for further use in molecular dynamics. Mutation of serine residues S268 and S311 to either glutamic acid (*E*) or phosphoserine were done with MODELLER. The initial structure of C-terminal peptide bound to PDZ domain was taken from the Coarse-Grained simulation using PLUM parameterization[65], where a position of the C-terminal peptide was sampled around the whole PDZ domain and most probable configuration was chosen as a starting structure.

We employed all-atom force field amber99sb-ILDN[66] with phosphoserine parameters[67] in explicit TIP3P[68] water model. Robert Best's correction for disordered proteins/regions was applied by scaling van der Waals interactions between water oxygen and protein with factor 1.1 (ref. [69]). All systems were solvated in ~10,000 water molecules placed in a cubic box, where the protein is surrounded by at least 1.3 nm from the box side. Protonation state of all residues was chosen according to neutral pH = 7 with histidine H324 neutral. If not stated otherwise, all proteins were capped at C-terminus by an acetyl group and at N-terminus by *N*-methyl to remove artificial charge introduced at protein ends.

All simulations were performed for at least 300 ns with 2 fs time step. The systems were kept in NPT ensemble at 309.15 K. Temperature was controlled using a velocity-rescaling thermostat[70] with a coupling constant of 0.1 ps applied separately to solvent and protein. The pressure was hold at 1 bar via an isotropic Parrinello–Rahman barostat[71,72], with coupling time 2 ps. Cut-off distance for direct space electrostatic and van der Waals interactions was set to 1 nm. Long-range electrostatic contribution was evaluated by using the particle-mesh Ewald summation method[73,74], with a maximal distance between FFT grid points of 0.12 nm and fourth interpolation order. All covalent bonds of hydrogen atoms were constrained with LINCS algorithm[75]. All systems were simulated using periodic boundary conditions.

The mean enthalpic energy contribution for each amino acid pair was calculated from the sum of the Lennard–Jones and Coulombic contributions for the whole residue including both backbone and side-chain atoms. Averages were taken from the whole 500 ns long production dynamics.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data Availability

Data supporting the findings of this manuscript are available from the corresponding author upon reasonable request. A reporting summary for this article is available as a Supplementary Information file. The source data underlying Figs. 1d, 2a, c, d, 3e, g–h, j, 4a, c, d, 5b–e, 6h, j, 7e, f, 8b–d and Supplementary Figs 1c, e, 2a, b, 4a–d, f, 7b are provided as a Source Data file. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[76] partner repository with the dataset identifier PXD013085. Figures were deposited in figshare depository possessing the identifier DOI 10.6084/m9.figshare.7856168

Received: 16 February 2018 Accepted: 20 March 2019
Published online: 18 April 2019

References

1. Clevers, H. Wnt/beta-catenin signaling in development and disease. *Cell* **127**, 469–480 (2006).
2. Gao, C. & Chen, Y. G. Dishevelled: the hub of Wnt signaling. *Cell Signal.* **22**, 717–727 (2010).
3. Wallingford, J. B. & Habas, R. The developmental biology of Dishevelled: an enigmatic protein governing cell fate and cell polarity. *Development* **132**, 4421–4436 (2005).
4. Mlodzik, M. The Dishevelled protein family: still rather a mystery after over 20 years of molecular studies. *Curr. Top. Dev. Biol.* **117**, 75–91 (2016).
5. Gentzel, M. & Schambony, A. Dishevelled paralogs in vertebrate development: redundant or distinct? *Front. Cell Dev. Biol.* **5**, 59 (2017).
6. Wynshaw-Boris, A. Dishevelled: in vivo roles of a multifunctional gene family during development. *Curr. Top. Dev. Biol.* **101**, 213–235 (2012).
7. Madrzak, J. et al. Ubiquitination of the Dishevelled DIX domain blocks its head-to-tail polymerization. *Nat. Commun.* **6**, 6718 (2015).
8. Cheyette, B. N. et al. Dapper, a Dishevelled-associated antagonist of beta-catenin and JNK signaling, is required for notochord formation. *Dev. Cell* **2**, 449–461 (2002).
9. Wong, H. C. et al. Structural basis of the recognition of the dishevelled DEP domain in the Wnt signaling pathway. *Nat. Struct. Biol.* **7**, 1178–1184 (2000).
10. Schwarz-Romond, T. et al. The DIX domain of Dishevelled confers Wnt signaling by dynamic polymerization. *Nat. Struct. Mol. Biol.* **14**, 484–492 (2007).
11. Fiedler, M., Mendoza-Topaz, C., Rutherford, T. J., Mieszczanek, J. & Bienz, M. Dishevelled interacts with the DIX domain polymerization interface of Axin to interfere with its function in down-regulating beta-catenin. *Proc. Natl Acad. Sci. USA* **108**, 1937–1942 (2011).
12. Schwarz-Romond, T., Metcalfe, C. & Bienz, M. Dynamic recruitment of axin by Dishevelled protein assemblies. *J. Cell Sci.* **120**, 2402–2412 (2007).
13. Bilic, J. et al. Wnt induces LRP6 signalosomes and promotes dishevelled-dependent LRP6 phosphorylation. *Science* **316**, 1619–1622 (2007).
14. Gammons, M. V., Renko, M., Johnson, C. M., Rutherford, T. J. & Bienz, M. Wnt signalosome assembly by DEP domain swapping of Dishevelled. *Mol. Cell* **64**, 92–104 (2016).
15. Smietana, K., Mateja, A., Krezel, A. & Otlewski, J. PDZ domain from Dishevelled—a specificity study. *Acta Biochim. Pol.* **58**, 243–249 (2011).
16. Lee, H. J., Shi, D. L. & Zheng, J. J. Conformational change of Dishevelled plays a key regulatory role in the Wnt signaling pathways. *eLife* **4**, e08142 (2015).
17. Qi, J. et al. Autoinhibition of Dishevelled protein regulated by its extreme C terminus plays a distinct role in Wnt/beta-catenin and Wnt/planar cell polarity (PCP) signaling pathways. *J. Biol. Chem.* **292**, 5898–5908 (2017).
18. Hoffmann, C. et al. A FlAsH-based FRET approach to determine G protein-coupled receptor activation in living cells. *Nat. Methods* **2**, 171–176 (2005).
19. Hoffmann, C. et al. Fluorescent labeling of tetracysteine-tagged proteins in intact cells. *Nat. Protoc.* **5**, 1666–1677 (2010).
20. Nuber, S. et al. beta-Arrestin biosensors reveal a rapid, receptor-dependent activation/deactivation cycle. *Nature* **531**, 661–664 (2016).
21. Xue, B., Dunbrack, R. L., Williams, R. W., Dunker, A. K. & Uversky, V. N. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta* **1804**, 996–1010 (2010).
22. Paclikova, P., Bernatik, O., Radaszkiewicz, T. W. & Bryja, V. N-terminal part of Dishevelled DEP domain is required for Wnt/beta-catenin signaling in mammalian cells. *Mol. Cell Biol.* https://doi.org/10.1128/MCB.00145-17 (2017).
23. Yanagawa, S., van Leeuwen, F., Wodarz, A., Klingensmith, J. & Nusse, R. The dishevelled protein is modified by wingless signaling in Drosophila. *Gene Dev.* **9**, 1087–1097 (1995).
24. Bryja, V., Schulte, G., Rawal, N., Grahn, A. & Arenas, E. Wnt-5a induces Dishevelled phosphorylation and dopaminergic differentiation via a CK1-dependent mechanism. *J. Cell Sci.* **120**, 586–595 (2007).
25. Rothbacher, U. et al. Dishevelled phosphorylation, subcellular localization and multimerization regulate its role in early embryogenesis. *EMBO J.* **19**, 1010–1022 (2000).
26. Gonzalez-Sancho, J. M. et al. Functional consequences of Wnt-induced Dishevelled 2 phosphorylation in canonical and noncanonical Wnt signaling. *J. Biol. Chem.* **288**, 9428–9437 (2013).
27. Bryja, V., Schulte, G. & Arenas, E. Wnt-3a utilizes a novel low dose and rapid pathway that does not require casein kinase 1-mediated phosphorylation of Dvl to activate beta-catenin. *Cell Signal.* **19**, 610–616 (2007).
28. Bernatik, O. et al. Sequential activation and inactivation of Dishevelled in the Wnt/beta-catenin pathway by casein kinases. *J. Biol. Chem.* **286**, 10396–10410 (2011).
29. Bernatik, O. et al. Functional analysis of dishevelled-3 phosphorylation identifies distinct mechanisms driven by casein kinase 1 and frizzled5. *J. Biol. Chem.* **289**, 23520–23533 (2014).
30. Kaucka, M. et al. Asymmetry of VANGL2 in migrating lymphocytes as a tool to monitor activity of the mammalian WNT/planar cell polarity pathway. *Cell Commun. Signal.* **13**, 2 (2015).
31. Cong, F., Schweizer, L. & Varmus, H. Casein kinase Iepsilon modulates the signaling specificities of dishevelled. *Mol. Cell Biol.* **24**, 2000–2011 (2004).
32. Xue, B., Dunker, A. K. & Uversky, V. N. Retro-MoRFs: identifying protein binding sites by normal and reverse alignment and intrinsic disorder prediction. *Int. J. Mol. Sci.* **11**, 3725–3747 (2010).
33. Foldynova-Trantirkova, S. et al. Breast cancer-specific mutations in CK1epsilon inhibit Wnt/beta-catenin and activate the Wnt/Rac1/JNK and NFAT pathways to decrease cell adhesion and promote cell migration. *Breast Cancer Res.* **12**, R30 (2010).
34. Luck, K., Charbonnier, S. & Trave, G. The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains. *FEBS Lett.* **586**, 2648–2661 (2012).
35. Mostarda, S., Gfeller, D. & Rao, F. Beyond the binding site: the role of the beta (2)-beta(3) loop and extra-domain structures in PDZ domains. *PLoS Comput. Biol.* **8**, e1002429 (2012).
36. Zhang, J. et al. Structural basis of beta-catenin recognition by Tax-interacting protein-1. *J. Mol. Biol.* **384**, 255–263 (2008).
37. Strakova, K. et al. Dishevelled enables casein kinase 1-mediated phosphorylation of Frizzled 6 required for cell membrane localization. *J. Biol. Chem.* **293**,18477–18493 (2018).
38. Jung, H. et al. Deubiquitination of Dishevelled by Usp14 is required for Wnt signaling. *Oncogenesis* **2**, e64 (2013).
39. Kishida, M. et al. Synergistic activation of the Wnt signaling pathway by Dvl and casein kinase Iepsilon. *J. Biol. Chem.* **276**, 33147–33155 (2001).
40. Rauch, J., Volinsky, N., Romano, D. & Kolch, W. The secret life of kinases: functions beyond catalysis. *Cell Commun. Signal.* **9**, 23 (2011).
41. Klein, T. J., Jenny, A., Djiane, A. & Mlodzik, M. CKIepsilon/discs overgrown promotes both Wnt-Fz/beta-catenin and Fz/PCP signaling in Drosophila. *Curr. Biol.* **16**, 1337–1343 (2006).
42. Wong, H. C. et al. Direct binding of the PDZ domain of Dishevelled to a conserved internal sequence in the C-terminal region of frizzled. *Mol. Cell* **12**, 1251–1260 (2003).
43. Punchihewa, C., Ferreira, A. M., Cassell, R., Rodrigues, P. & Fujii, N. Sequence requirement and subtype specificity in the high-affinity interaction between human frizzled and dishevelled proteins. *Protein Sci.* **18**, 994–1002 (2009).
44. Witte, F. et al. Negative regulation of Wnt signaling mediated by CK1-phosphorylated Dishevelled via Ror2. *FASEB J.* **24**, 2417–2426 (2010).
45. Funato, Y. et al. Nucleoredoxin sustains Wnt/beta-catenin signaling by retaining a pool of inactive dishevelled protein. *Curr. Biol.* **20**, 1945–1952 (2010).
46. Funato, Y., Michiue, T., Asashima, M. & Miki, H. The thioredoxin-related redox-regulating protein nucleoredoxin inhibits Wnt-beta-catenin signalling through dishevelled. *Nat. Cell Biol.* **8**, 501–U135 (2006).
47. Angers, S. et al. The KLHL12-Cullin-3 ubiquitin ligase negatively regulates the Wnt-beta-catenin pathway by targeting Dishevelled for degradation. *Nat. Cell Biol.* **8**, 348–357 (2006).
48. McKay, R. M., Peters, J. M. & Graff, J. M. The casein kinase I family in Wnt signaling. *Dev. Biol.* **235**, 388–396 (2001).
49. Ran, F. A. et al. Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
50. Chu, V. T. et al. Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nat. Biotechnol.* **33**, 543–548 (2015).
51. Bryja, V. et al. Increased apoptosis in differentiating p27-deficient mouse embryonic stem cells. *Cell. Mol. Life Sci.* **61**, 1384–1400 (2004).
52. Cervenka, I. et al. Dishevelled is a NEK2 kinase substrate controlling dynamics of centrosomal linker proteins. *Proc. Natl Acad. Sci. USA* **113**, 9304–9309 (2016).
53. Nieuwkoop, P. D. & Faber Garland, J. *Normal table of Xenopus laevis* (Garland Publishing Inc, New York, 1994).
54. Gentzel, M., Schille, C., Rauschenberger, V. & Schambony, A. Distinct functionality of dishevelled isoforms on Ca2+/calmodulin-dependent protein kinase 2 (CamKII) in Xenopus gastrulation. *Mol. Biol. Cell* **26**, 966–977 (2015).
55. Bryja, V. et al. Beta-arrestin and casein kinase 1/2 define distinct branches of non-canonical WNT signalling pathways. *EMBO Rep.* **9**, 1244–1250 (2008).
56. Vilardaga, J. P., Bunemann, M., Krasel, C., Castro, M. & Lohse, M. J. Measurement of the millisecond activation switch of G protein-coupled receptors in living cells. *Nat. Biotechnol.* **21**, 807–812 (2003).
57. Jost, C. A., Reither, G., Hoffmann, C. & Schultz, C. Contribution of fluorophores to protein kinase C FRET probe performance. *Chembiochem* **9**, 1379–1384 (2008).

58. Stejskal, K., Potesil, D. & Zdrahal, Z. Suppression of peptide sample losses in autosampler vials. *J. Proteome Res.* **12**, 3057–3062 (2013).
59. Stanek, J., Augustyniak, R. & Kozminski, W. Suppression of sampling artefacts in high-resolution four-dimensional NMR spectra using signal separation algorithm. *J. Magn. Reson.* **214**, 91–102 (2012).
60. Evangelidis, T. et al. Automated NMR resonance assignments and structure determination using a minimal set of 4D spectra. *Nat. Commun.* **9**, 384 (2018).
61. Pronk, S. et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–854 (2013).
62. Sali, A. & Blundell, T. L. Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
63. Fiser, A. & Sali, A. MODELLER: Generation and refinement of homology-based protein structure models. *Methods Enzymol.* **374**, 461–491 (2003).
64. Fiser, A., Do, R. K. G. & Sali, A. Modeling of loops in protein structures. *Protein Sci.* **9**, 1753–1773 (2000).
65. Bereau, T. & Deserno, M. Generic coarse-grained model for protein folding and aggregation. *J. Chem. Phys.* **130**, 235106 (2009).
66. Lindorff-Larsen, K. et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–1958 (2010).
67. Homeyer, N., Horn, A. H., Lanig, H. & Sticht, H. AMBER force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine. *J. Mol. Model.* **12**, 281–289 (2006).
68. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
69. Best, R. B., Zheng, W. W. & Mittal, J. Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.* **10**, 5113–5124 (2014).
70. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
71. Parrinello, M. & Rahman, A. Polymorphic transitions in single-crystals—a new molecular-dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
72. Nose, S. & Klein, M. L. Constant pressure molecular-dynamics for molecular-systems. *Mol. Phys.* **50**, 1055–1076 (1983).
73. Darden, T., York, D. & Pedersen, L. Particle Mesh Ewald—an N.Log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
74. Essmann, U. et al. A smooth particle Mesh Ewald Method. *J. Chem. Phys.* **103**, 8577–8593 (1995).
75. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).
76. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).

Acknowledgements

We would like to thank Lucie Nesvadbová, Lenka Bryjová, Pavlína Žofka Mrhálková, and Naďa Bílá for excellent assistance, and the members of the Bryja Lab for helpful discussions. Also, we would like to thank Petra Ovesná for help with the statistics in Fig. 8. Special thanks go to Kathyann Lee for the help with English Language editing and review. The work was supported by Czech Science Foundation (project no. GA15-21789S, GA18-17658S, and GA19-26041x), Grant Agency of Masaryk University (MUNI/G/1100/2016, MUNI/G/0739/2017, MUNI/E/0533/2018). J.H. specially thanks for the support of the EMBO Short-Term Fellowship (no. ASTF 687-2016). Furthermore, this research was supported by project CEITEC 2020 (no. LQ1601) with financial contribution from the MEYS CR and National Programme for Sustainability II to R.V., M.J., K.T., L.T., K.H. and J.R. and by the; project SYMBIT (CZ.02.1.01/0.0/0.0/15_003/0000477) funded by the European Regional Development Fund and by the MEYS CR; CIISB research infrastructure project LM2015043 funded by MEYS CR is gratefully acknowledged for the financial support of the measurements at CEITEC Josef Dadok National NMR Centre and Proteomics Core Facility. J.R. thanks for support of Ministry of Health of the Czech Republic project (no. 15-34405A). M.C.A.C. and C. Hoffmann were supported by the ITN WntsApp (grant no. 608180; www.wntsapp.eu) and the Deutsche Forschungs-gemeinschaft (DFG, German Research Foundation)—TRR166: project C2. Computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085 provided under the program "Projects of Large Research, Development, and Innovations Infrastructures". This work was supported from European Structural and Investment Funds, Operational Programme Research, Development and Education – "Preclinical Progression of New Organic Compounds with Targeted Biological Activity" (Preclinprogress) -CZ.02.1.01/0.0/0.0/16_025/0007381. Additional computational resources were obtained from 'IT4 Innovations National Supercomputing Center -- LM2015070' project supported by MEYS CR from the Large Infrastructures for Research, Experimental Development and Innovations. C. Holler and A.S. were supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), grant SCHA965/6-2.

Author contributions

J.H., K.T., R.V., C. Hoffmann., and V.B. designed the research. J.H. performed all the research and data visualization except those mentioned below. M.C.A.C. contributed to the FlAsH-based FRET approach; M.J. and R.V. performed the in silico simulations and corresponding data visualization; J.K. and K.T. performed the NMR and SPR analyses and corresponding data visualization; C. Holler and A.S. performed the in vivo analyses in *Xenopus* and corresponding data visualization; K.H., O.B., and Z.Z. performed the MS/MS-based analysis; K.G. contributed to the IP and WB approach; T.G. and T.W.R. cloned the cell lines (CK1ε$^{-/-}$ and ECFP-DVL3 FIII cells); M.K. cloned and purified the recombinant DVL3 protein; L.T. and J.R. contributed to the peptide array approach; Z.D. and A.I.F.-L. synthesized the peptides for the FA approach; J.H., J.K., M.J., K.H., Z.Z., R.V., A.S., K.T., C. Hoffmann., and V.B. analyzed data; J.H., R.V., K.T., and V.B. wrote the original manuscript; J.H. and V.B. revised and edited the paper.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-09651-7.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Journal peer review information: *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-23314-6 OPEN

Distinct EH domains of the endocytic TPLATE complex confer lipid and protein binding

Klaas Yperman[1,2], Anna C. Papageorgiou[3,10], Romain Merceron[4,5,10], Steven De Munck[4,5], Yehudi Bloch[4,5], Dominique Eeckhout[1,2], Qihang Jiang[1,2], Pieter Tack[6], Rosa Grigoryan[7], Thomas Evangelidis[3], Jelle Van Leene[1,2], Laszlo Vincze[6], Peter Vandenabeele[6,8], Frank Vanhaecke[7], Martin Potocký[9], Geert De Jaeger[1,2], Savvas N. Savvides[4,5]✉, Konstantinos Tripsianes[3]✉, Roman Pleskot[1,2,9]✉ & Daniel Van Damme[1,2]✉

Clathrin-mediated endocytosis (CME) is the gatekeeper of the plasma membrane. In contrast to animals and yeasts, CME in plants depends on the TPLATE complex (TPC), an evolutionary ancient adaptor complex. However, the mechanistic contribution of the individual TPC subunits to plant CME remains elusive. In this study, we used a multidisciplinary approach to elucidate the structural and functional roles of the evolutionary conserved N-terminal Eps15 homology (EH) domains of the TPC subunit AtEH1/Pan1. By integrating high-resolution structural information obtained by X-ray crystallography and NMR spectroscopy with all-atom molecular dynamics simulations, we provide structural insight into the function of both EH domains. Both domains bind phosphatidic acid with a different strength, and only the second domain binds phosphatidylinositol 4,5-bisphosphate. Unbiased peptidome profiling by mass-spectrometry revealed that the first EH domain preferentially interacts with the double N-terminal NPF motif of a previously unidentified TPC interactor, the integral membrane protein Secretory Carrier Membrane Protein 5 (SCAMP5). Furthermore, we show that AtEH/Pan1 proteins control the internalization of SCAMP5 via this double NPF peptide interaction motif. Collectively, our structural and functional studies reveal distinct but complementary roles of the EH domains of AtEH/Pan1 in plant CME and connect the internalization of SCAMP5 to the TPLATE complex.

[1] Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium. [2] VIB Center for Plant Systems Biology, Ghent, Belgium. [3] CEITEC-Central European Institute of Technology, Masaryk University, Brno, Czech Republic. [4] Department of Biochemistry and Microbiology, Ghent University, Ghent, Belgium. [5] VIB Center for Inflammation Research, Ghent, Belgium. [6] Department of Chemistry, X-ray Microspectroscopy and Imaging – XMI Research Unit, Ghent University, Ghent, Belgium. [7] Department of Chemistry, Atomic & Mass Spectrometry – A&MS Research Unit, Ghent University, Ghent, Belgium. [8] Archaeometry Research Group, Department of Archaeology, Ghent University, Ghent, Belgium. [9] Institute of Experimental Botany, Academy of Sciences of the Czech Republic, Prague 6, Czech Republic. [10] These authors contributed equally: Anna C. Papageorgiou, Romain Merceron. ✉email: savvas.savvides@irc.vib-ugent.be; kostas.tripsianes@ceitec.muni.cz; pleskot@ueb.cas.cz; daniel.vandamme@psb.vib-ugent.be

Internalization of membrane proteins is of crucial importance for cell survival as it allows to quickly react to changing environmental conditions. The residence time of integral membrane proteins at the plasma membrane is controlled by internalization signals that are recognized by adaptor protein complexes, which mediate a process named clathrin-mediated endocytosis (CME). The start of CME is marked by an enrichment of cargo proteins and negatively charged phospholipids[1]. In the next stage, adaptor proteins and clathrin are recruited in a timed fashion and accumulate at the site of endocytosis[2]. In the later stages, accessory proteins and additional clathrin molecules are recruited followed by scission of the formed vesicle from the plasma membrane. The main drivers during this process are low affinity protein–protein and protein–lipids interactions, which enable this dynamic process of assembly and disassembly of clathrin-coated vesicles.

In plants, two adaptor complexes play a role during the initiation phase of endocytosis, the Adaptor Protein-2 complex (AP-2) and the TPLATE complex (TPC)[3,4]. AP-2 and TPC likely have independent but also complementary roles in CME[3]. Both protein complexes have a core-complex of four subunit[5], which in the case of TPC is associated with four additional subunits (TWD40-1, TWD40-2, AtEH1/Pan1, and AtEH2/Pan1)[6]. The AtEH/Pan1 proteins are more loosely associated with TPC as they for example do not associate with the other complex subunits when a truncated TML subunit forces TPC into the cytoplasm[3] and they do not co-purify with the complex in *Dictyostelium*[5]. Recent evidence suggests however a similar arrival time of all TPC subunits at the plasma membrane at the onset of endocytosis, preceding clathrin arrival[3,7,8].

The AtEH/Pan1 proteins are the plant homologs of yeast Pan1p, which is known for its role as an activator of ARP2/3-dependent actin dynamics during endocytosis[9–11]. Pan1p was also recently shown to be part of a phosphorylation-dependent mechanism connecting endocytic vesicles, endosomal compartments, and actin dynamics in budding yeast[12]. The Arabidopsis AtEH/Pan1 proteins were recently shown to mediate actin-dependent autophagy in plants[13]. AtEH/Pan1 and Pan1p proteins are both hallmarked by the presence of two EH domains at their N-terminus[3,13].

In animals and yeast, EH domains have been characterized in great detail due to their presence in crucial endocytic proteins like Eps15, REPS1, EHD1, etc.[14–16] SMART[17] and Prosite[18] analysis identified only six EH domains in Arabidopsis compared to eighteen in humans. In Arabidopsis, EH domains are present in the endocytic recycling regulators EHD1 and EHD2, each having one EH domain[19] and in the TPC subunits AtEH1/Pan1 and AtEH2/Pan1, each having two EH domains. Plant EHD1 and EHD2 proteins have been characterized as homologs of human EHD proteins and were suggested to play a regulatory role during plant CME, plant defense, and salt stress[19–21]. In contrast to the essential function of all tested TPC subunits, silencing of EHD1/2 does not result in severely aberrant phenotypes, indicating redundancy or a more specialized function[3,13,19]. In contrast to the single EH domain in EHD proteins, AtEH/Pan1 proteins harbor two EH domains and we asked if both EH domains serve as independent functional modules. We, therefore, set out to characterize the function of the EH domains of the AtEH/Pan1 proteins in Arabidopsis. In this study, we used a multidisciplinary approach to perform a structural and functional side-by-side comparison of both EH domains of AtEH1/Pan1.

Results and discussion

Structural characterization of both EH domains of AtEH1/Pan1 reveals a common fold. AtEH/Pan1 proteins are highly unstructured but three domains can be identified; a coiled-coil domain implicated in dimerization[22] and two N-terminal Eps15 homology (EH) domains (Fig. 1a, Supplementary Fig. 1). Comparing the EH domains within each AtEH/Pan1 protein shows low sequence identity which is in contrast to the high sequence identity when comparing the particular EH domain between AtEH1/Pan1 and AtEH2/Pan1 (Fig. 1a). Therefore, we decided to focus on both EH domains of AtEH1/Pan1 as representatives, which we hereafter name EH1.1 and EH1.2. To structurally characterize both EH domains of AtEH1/Pan1, we expressed recombinant proteins in *E.coli* and purified highly monodisperse samples for X-ray crystallography and NMR (Supplementary Fig. 1). Crystals of EH1.1 diffracted synchrotron X-rays to 1.55 Å resolution and enabled structure determination via molecular replacement (PDB:6YIG) (Supplementary Table 1). EH1.1 consists of two EF-hands connected by two short antiparallel β-sheets (Fig. 1b). Anomalous scattering supports a calcium ion bound in a pentagonal-bipyramidal geometry in the loop of the first EF-hand (Fig. 1d; Supplementary Fig. 1). No anomalous scattering signal was detected in the second EF-hand loop, but electron density consistent with the coordination of a sodium ion was present (Fig. 1c). Attempts to crystallize EH1.2 were unsuccessful, therefore, we obtained structural insights for both EH1.1 (PDB: 6YEU) and EH1.2 (PDB: 6YET), in solution, by NMR Spectroscopy (Supplementary Fig. 1, Supplementary Table 2)[23]. With respect to EH1.1, the NMR and X-ray structures agree very well with a RMSD of 1 Å. When comparing the NMR structures of EH1.1 and EH1.2, the RMSD between both domains is 2.6 Å. In general, the EH1.1 and EH1.2 domains both have a typical EH domain fold with a very conserved hydrophobic core. The major difference between the two domains is at the interaction interface of the N- and C-terminal ends. The hydrophobic core of EH1.1 is shielded by the proline-rich C-terminal loop, while in EH1.2, the proline-rich N-terminal loop takes over this function.

The second EH domain coordinates two calcium ions. We were intrigued by the fact that, in contrast to EH1.1, EH1.2 contains two possible calcium-binding sites. This is manifested by a classical calcium-binding motif (DxDxDxxxxxxE) in the second loop of EH1.2 and a second possible coordination motif in the first loop, where the final glutamate, in the canonical calcium-ligation cassette, is substituted by glutamine in several species including Arabidopsis (Supplementary Fig. 1). Experimental assessment by two independent methods, total reflection X-ray fluorescence (TXRF) and inductively coupled plasma-mass spectrometry (ICP-MS), confirmed the presence of two calcium ions in the Arabidopsis EH1.2 domain (Fig. 1h, i). To our knowledge, the ability of EH domains to coordinate two calcium ions has not been described before. To obtain a possible coordination scheme for Arabidopsis, we performed an extensive all-atom molecular dynamics simulation (3 µs, CHARMM force-field) (Fig. 1e–g, Supplementary Fig. 2, Supplementary Data 6). The all-atom molecular dynamics model, within its time and forcefield limitations, suggests that the first aspartate and the presence of an extra water molecule compensate for the incomplete calcium-binding motif in the first loop of EH1.2 and functionally mimics the role of the canonical glutamate in the calcium-binding motif (Fig. 1g, Supplementary Fig. 2). Restoring the first EF-hand loop to a canonical EF-hand (Q382E), in an all-atom molecular dynamics simulation, resulted in a classical arrangement where glutamate coordinates calcium in a bidentate fashion (Supplementary Fig. 2, Supplementary Data 6). We confirmed by NMR that EH1.2 is indeed more sensitive to precipitation upon calcium chelation compared to EH1.1 (Fig. 1j). This is consistent with our findings that EH1.2 coordinates two calcium ions.

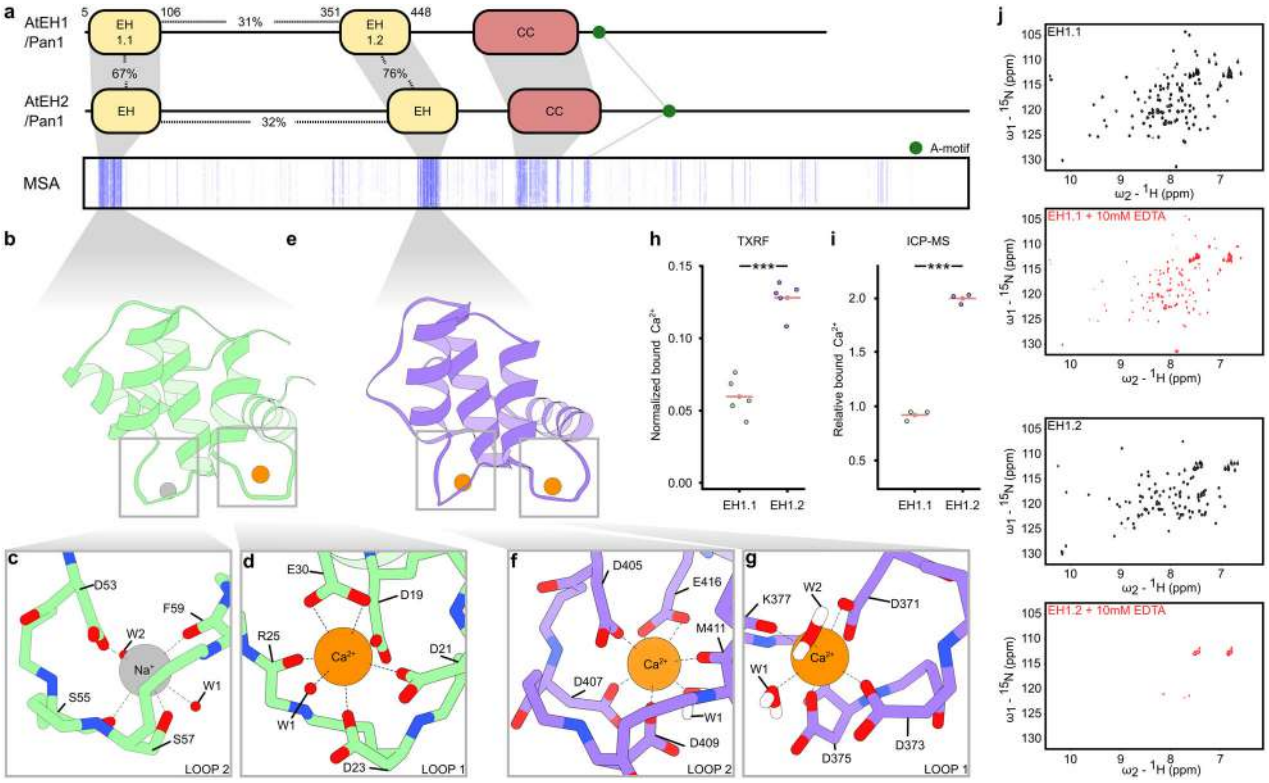


Fig. 1 EH domains of AtEH1/Pan1 differ in their Ca^{2+}-binding capacities. a Domain organization of AtEH1/Pan1 and AtEH2/Pan1. Both proteins contain two Eps15 homology domains (EH), a coiled-coil domain (CC), and an acidic (A)-motif. A schematic representation of a multiple sequence alignment (MSA), shows strong conservation of the EH domains (blue lines) across the plant kingdom. Percentages indicate the relative number of identical amino acids. **b**–**g** Cartoon representation of the X-ray structure of EH1.1 and NMR/all-atom molecular dynamics structure of EH1.2. Ions are shown as orange (Ca^{2+}) or grey (Na^{+}) spheres. Insets show the ion coordination in each EF-hand loop. Ca^{2+} coordinating residues and water molecules (W) are indicated in (**c**, **d**) and (**f**, **g**). **h** Total reflection X-ray fluorescence (TXRF) intensities of Ca^{2+} normalized to Cl^{-} of samples containing 2 mM of each EH domain in the presence of 0.5 mM free Ca^{2+}. The mean ($n = 5$) is indicated as a pink line. Statistical analysis was performed using a two-tailed Welch's *t*-test ($p = 1.056 \times 10^{-4}$). **i** Ca^{2+} concentration relative to the amount of protein as measured by inductively coupled plasma mass spectrometry (ICP-MS) after subtraction of the amount of free Ca^{2+} in the medium. The mean ($n = 3$) is indicated as a pink line. Statistical analysis was performed using a two-tailed Welch's *t*-test ($p = 1.046 \times 10^{-5}$). **j** ^{15}N-^{1}H HSQC spectra of each EH domain before (black) and after (red) Ca^{2+} chelation by 10 mM EDTA.

AtEH1/Pan1 EH domains differently interact with charged lipids. EH domains have been proposed to act as a protein interaction hub and/or as a lipid-binding module[24,25]. To unravel the function of the EH domains of AtEH1/Pan1, we tested both domains for their ability to bind peptide motifs and membranes, in a pairwise manner. AtEH/Pan1 proteins mainly function at the negatively charged plasma membrane or the ER-PM contact sites[3,13]. To mimic a negatively charged plasma membrane environment in vitro we performed liposome-binding experiments using an equimolar mixture of phosphatidylethanolamine and phosphatidylcholine (PC) with or without 10% phosphatidylinositol 4,5-bisphosphate (PIP_2). Only EH1.2 substantially bound to PIP_2 enriched liposomes (Fig. 2a, b). Next, we tested the binding of EH1.2 to monophosphate phosphoinositides. We observed only a very weak interaction with phosphatidylinositol 3-phosphate (PI3P) and phosphatidylinositol 4-phosphate (PI4P) compared to PIP_2 enriched liposomes (Supplementary Fig. 3). Given the fact that, on the one hand, calcium is needed for the fold of EH1.2 and, on the other hand, that lipid-EH1.2 interactions are electrostatically-driven, we tested the effect of PIP_2 binding by the EH1.2 domain at different calcium concentrations. We observed increased binding of EH1.2 at lower calcium concentrations (Supplementary Fig. 3). Our results are in line with prior studies showing that high calcium concentrations block the accessibility of charged lipids by shielding and rearranging the lipid headgroups[26].

To understand the difference in the PIP_2-binding of EH1.1 and EH1.2, we compared structural features of both AtEH1/Pan1 EH domains with the EH domains of Human Eps15 (EH2, PDB:1F8H) and EHD1 (EH1, PDB:2KSP), both known for their ability to interact with phosphoinositides[24]. The lipid interacting residues in the EH domains of EHD1 and Eps15 are structurally conserved in EH1.2 (R384, R391, and K398), whereas no lysine or arginine residues are present at those positions in EH1.1 (Supplementary Fig. 3). We generated a triple mutant of the EH1.2 domain (i.e. R384, R391, and K398 were mutated to glutamate, EH1.2 R/K>E) and tested PIP_2 binding (Supplementary Fig. 3). The obtained mutant protein was severely impaired in its interaction with the liposomes containing 10% PIP_2, further supporting that the positively charged residues on the surface are the determinants of the interaction with the PIP_2 molecules (Supplementary Fig. 3).

To show whether the EH domains of AtEH1/Pan1 interact with membranes *in planta*, we prepared them as a triple repeat of each EH domain fused C-terminally with GFP, allowing for an avidity effect. We expressed the three-tandem domains transiently in *N. benthamiana* together with cyclin D. Overexpression of the cyclin D protein triggers cell division in leaves of *N. benthamiana* and thus enables to monitor different stages of cytokinesis in this otherwise differentiated model system[27]. Both three-tandem domains localized to the early cell plate (Fig. 2c), which is the most negatively charged membrane compartment in

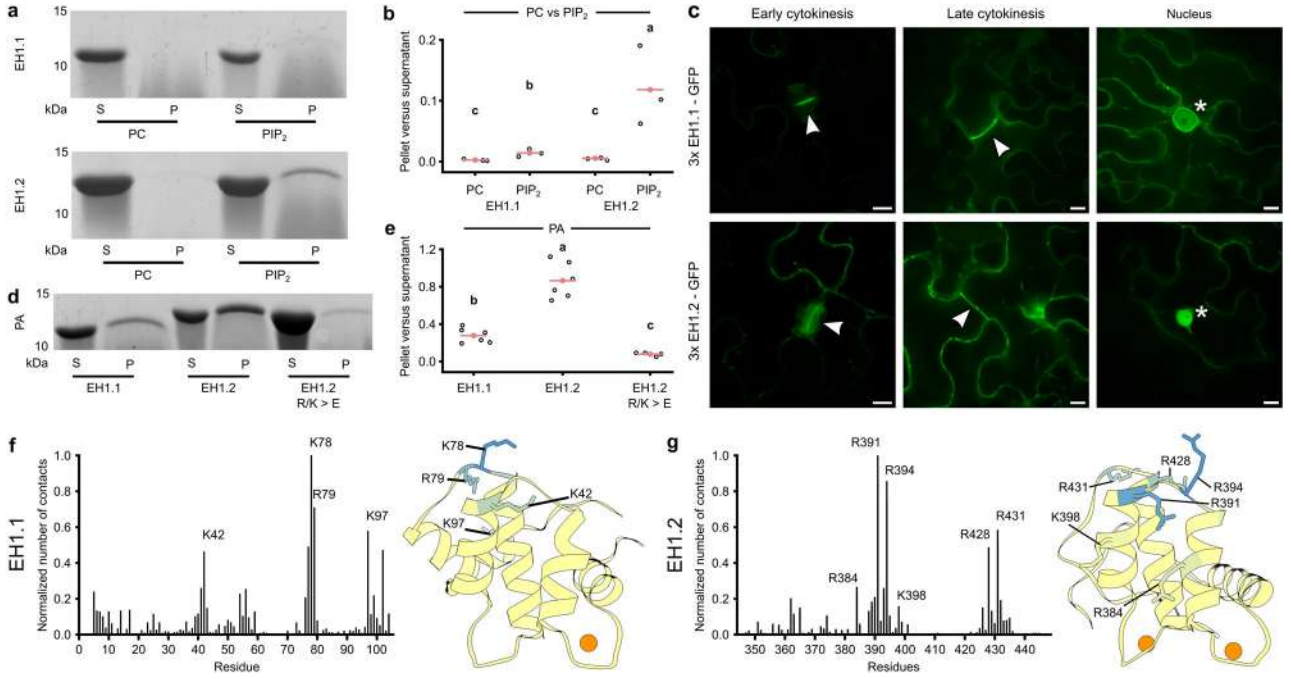


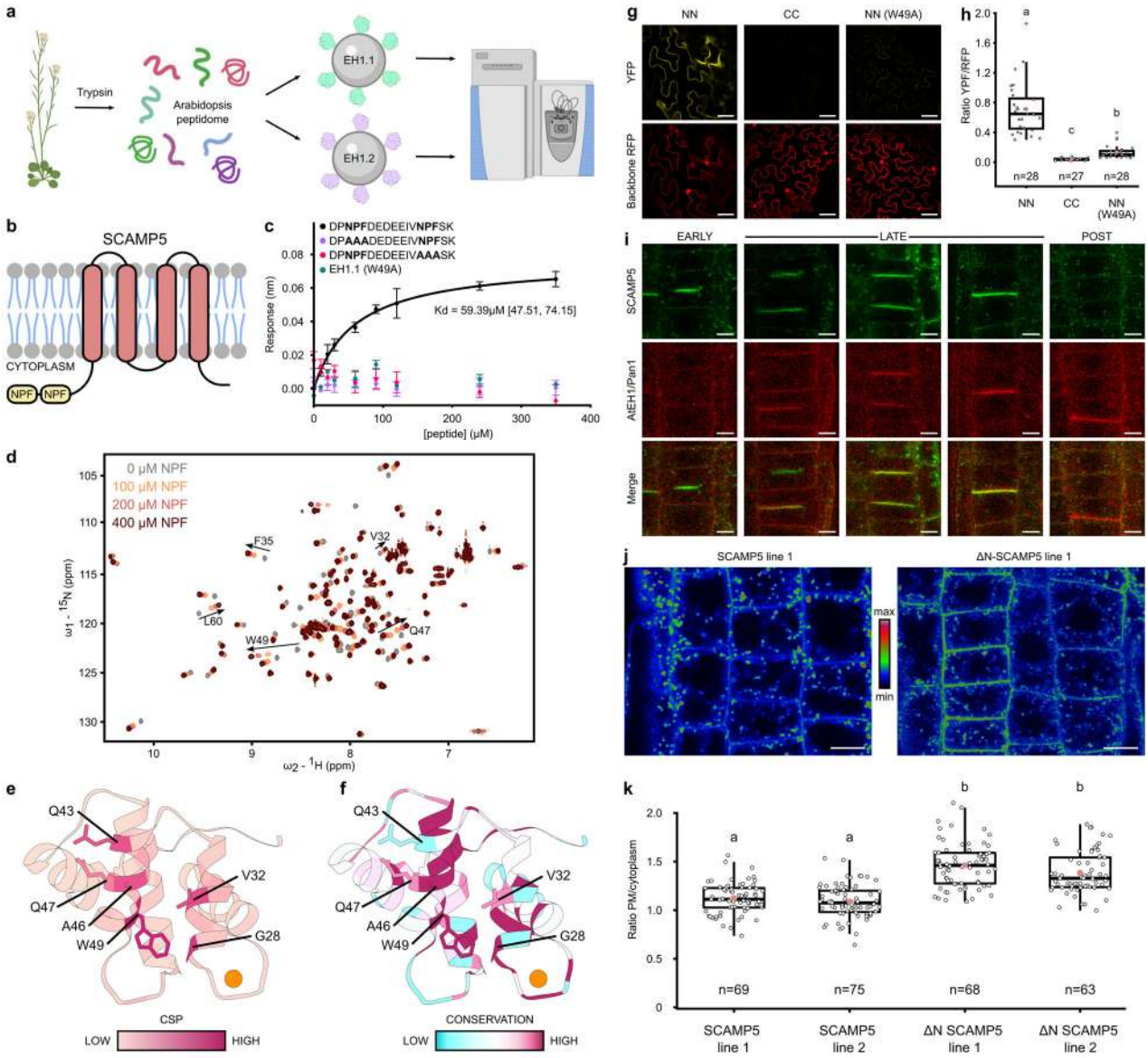
Fig. 2 Differential binding of EH domains of AtEH1/Pan1 to anionic phospholipids. a Coomassie-blue stained SDS-PAGE analysis of liposome binding comparing the binding of EH domains between PC (an equimolar mixture of PE and PC) and 10% PIP_2 containing liposomes in the presence of 10 µM Ca^{2+}. S = supernatant, P = pellet. **b** Quantification of lipid binding (ratio of cumulative grey values between pellet and supernatant) as shown in (**a**). Different letters indicate significant differences between samples using a one-sided Kruskal–Wallis test ($p \leq 0.05$). $n = 3$ independent experiments. **c** Localization of triple EH domains fused C-terminally to GFP in *N. benthamiana* epidermal leaf cells. Cells were triggered to divide by overexpression of the Cyclin D protein. The early and late stages of cell division are depicted (arrowheads). In contrast to EH1.2, EH1.1 is retained more prominent at the late cell plate (arrowhead) and also labels the nuclear envelope (asterisk). Scale bar indicates 10 µm. The experiment was repeated twice independently with similar results. **d** Coomassie-blue stained SDS-PAGE analysis of PA liposome binding of both EH domains as well EH1.2 mutated in its predicted PIP_2-binding site (EH1.2 R/K>E). S = supernatant, P = pellet. **e** Quantification of PA lipid binding (pellet versus supernatant) as shown in (**d**). Different letters indicate significant differences between samples using a one-sided Kruskal–Wallis test ($p \leq 0.05$). $n = 4$ or 5 independent experiments. **f**, **g** CG-MD simulations of EH1.1 (**f**) and EH1.2 (**g**) with a lipid bilayer containing 20% PA. On the left, the mean number of contacts with PA for each domain are shown. The contacts were defined as the number of PA phosphate groups within 0.8 nm of protein atoms calculated through the whole simulation time and averaged over all CG-MD replicas. On the right, a cartoon representation of each EH domain is shown. The color gradient (yellow to blue) indicates the extent of interaction with PA. Residues with the highest contact number are shown as sticks. Ca^{2+} is shown as an orange sphere.

plant cells[28]. In contrast to EH1.2, EH1.1 remained more prominently present at the cell plate in later stages of cytokinesis and also labeled the nuclear envelope (Fig. 2c). The cell plate is devoid of PIP_2, but enriched in phosphatidic acid (PA) from very early on[29]. Therefore, we tested a possible direct PA interaction with both EH domains by liposome-binding assays. Consistently with the membrane localization of EH domains *in planta*, we found that both domains directly interact with the liposomes containing 20% PA, with EH1.2 binding PA more efficiently than EH1.1 (Fig. 2d, e).

In order to elucidate the molecular mechanism of the interaction between PA and the EH domains of AtEH1/Pan1, we employed extensive coarse-grained molecular dynamics (CG-MD) simulations. During multiple 1 µs-simulations (in total 35), both domains interacted with the lipid bilayer containing 20% PA (Supplementary Fig. 4). However, we found that the membrane-binding mode of each EH domain substantially differed from each other (Fig. 2f-g). On the one hand, EH1.1 coordinates the PA molecules by three main regions (residues around K42, residues around K78, and residues located at the C-terminus around K97). On the other hand, EH1.2 interacts with the PA molecules predominantly by the region forming the PIP_2 binding site (R384, R391, and K398), but additionally, the second region located close to the protein C-terminus (R428 and R431) significantly contributes to the PA coordination. This was corroborated by the liposome-binding experiment with the triple mutant of EH1.2 in the PIP_2 binding site (EH1.2 R/K>E), which showed significantly reduced, but not completely abolished, binding to the PA (Fig. 2d, e). The observed residual PA-binding of the EH1.2 triple mutant is then likely mediated via a second motif, located C-terminally (Fig. 2g). Our data also suggest that, in contrast to EH1.1, EH1.2 coordinates a higher number of PA molecules (Supplementary Fig. 4). The above results are consistent with our biochemical data showing that EH1.2 interacts more efficiently with PA-containing liposomes than EH1.1 (Fig. 2d, e).

A recent report showed that PA significantly contributes to the electrostatic signature of the plant plasma membrane[29] and several plant proteins involved in CME, such as the ANTH domain-containing proteins AP180 and ECA1, were described to interact with PA[30–32]. Our data, showing that the EH domains of the AtEH/Pan1 proteins, which are plant-specific subunits of the TPC[6], also interact with PA thus suggest a more general role of PA in plant CME.

The first EH domain recognizes a previously unidentified endocytic transport motif in plants. To identify protein interaction partners of the AtEH1/Pan1 EH domains, we sought to discover interaction motifs. To this end, we digested Arabidopsis seedling proteome and incubated the peptide mix with each EH domain or with GFP as a control (Fig. 3a). Comparative mass spectrometry identified bound peptides. The N-terminal double



NPF peptide of SCAMP5 was identified as a significant interactor of EH1.1 (Fig. 3b, Supplementary Data 3). This peptide also showed the highest fold change compared to EH1.2 and GFP. Secretory Carrier Membrane Protein 5 (SCAMP5) is part of a five-membered protein family in Arabidopsis. SCAMP proteins were first characterized in mammals for their role in endocytosis[33]. Current insights in the animal and plant field show a broader function of SCAMPs in plasma membrane phase separation, cell plate formation, and pathogen-induced stomatal closure[34–36]. SCAMP5 was also previously identified to reside in close proximity to TPC by proximity labeling[37]. No peptides derived from integral membrane proteins were identified via SMART[17] and Uniprot database[38] searches among the few specific interactors of EH1.2.

Bio-Layer Interferometry (BLI) confirmed the interaction between EH1.1 and a peptide derived from the N-terminus of SCAMP5 with a binding affinity (K_d) of 59 μM via steady-state kinetics (Fig. 3c). Both NPF stretches of the N-terminus of SCAMP5 are required to bind EH1.1, as no binding affinity for mutated NPF to AAA mutants could be determined (Fig. 3c). The interaction between the double NPF motif is much stronger for EH1.1 ($K_{d,}^{app} \sim 33$ μM) compared to EH1.2 ($K_{d,}^{app} \sim 190$ μM) as shown by NMR titration experiments (Fig. 3d, Supplementary Fig. 5). NMR peptide titration experiments revealed the binding sites of the peptide on both EH domains (Fig. 3d, Supplementary Fig. 5). Published structures of EH domains interacting with NPF-motifs show the importance of the central tryptophan and its surrounding hydrophobic residues, in agreement with the largest chemical shift perturbations observed in EH1.1[14,15,39] (Fig. 3d, e). Comparison of chemical shift perturbation of EH1.1 with evolutionary conservation across a variety of plant species showed that most of the residues responsible for the NPF motif binding are strongly conserved (Fig. 3f). Mutating the conserved tryptophan to alanine (W49A) in EH1.1, and testing its binding capacity by BLI confirmed its essential role. No binding was observed for the mutated EH1.1 domain (Fig. 3c). We conclude that the N-terminal double NPF motif of SCAMP5 interacts preferentially with EH1.1, mediated via a hydrophobic interaction with the conserved tryptophan residue. The NPF-binding pocket in EH1.1, as identified by NMR chemical shift mapping, has been described before in other EH domains and a similar contribution of the residues surrounding the conserved tryptophan was shown[14,15,39]. In the identified SCAMP5 NPF motif, not one but two NPF motifs are present and separated by an acidic region.

Fig. 3 The first EH domain of AtEH1/Pan1 interacts with the N-terminal double NPF motif of SCAMP5. a Scheme of the peptidome profiling experiment. **b** Graphical representation of SCAMP5, the double N-terminal NPF motif is indicated in yellow. **c** BLI steady-state kinetics of the binding of EH1.1 of AtEH1/Pan1 and NPF peptides with or without mutations. WT Protein but not the W49A mutant binds the double NPF peptide with a measurable affinity. Mutation of any of the NPF motifs abrogates binding. Data are presented as mean values ± SD. $n = 3$ independent experiments. **d** ^{1}H-^{15}N-HSQC spectra of EH1.1 titrated with increasing amounts of the SCAMP5 double NPF peptide (grey to red). Peak trajectories of selected residues are indicated by arrows. Their weighted chemical shift perturbations were used to obtain binding isotherms and derive an apparent dissociation constant of the interaction (Supplementary Fig. 5). **e** Cartoon representation of EH1.1 (6-105) colored with a gradient (pink to purple) indicate the extent of chemical shift perturbations induced by the SCAMP5 double NPF peptide binding. Residues showing large chemical shift perturbations (>0.3 ppm) are shown as sticks. **f** Similar as in panel d but the EH1.1 structure is colored according to ConSurf colors denoting evolutionary conservation (blue to white to purple). Most residues affected by peptide binding are well-conserved. **g** Ratiometric bimolecular fluorescence complementation (rBiFC) analysis showing interaction between AtEH1/Pan1 and SCAMP5. Both proteins specifically interact when tagged at the N-terminus. Mutating the SCAMP5-binding site in the first EH domain of AtEH1/Pan1 (W49A) drastically reduces the interaction. Scale bar is 50 µm. **h** Quantification of the YFP/RFP fluorescence ratios from the experiment in (**g**). The black lines represent the median and the red circles represent the mean. The amount of quantified cells is indicated below each boxplot. Letters above the plots indicate statistically significant differences analyzed by one-sided Welch's ANOVA post hoc pairwise comparison was performed with the package multcomp utilizing the Tukey contrasts ($p \leq 0.001$). The boxplot extends from the 25th to 75th percentiles. The line inside the box marks the median. The whiskers go down and up to the 95% percentile. **i** Co-localization of SCAMP5-GFP and AtEH1/Pan1-mRuby3 at the plasma membrane and the cell plate in Arabidopsis root cells. Cells in different phases of cytokinesis (early, late and post) are depicted. SCAMP5 recruitment to the cell plate precedes AtEH1/Pan1 whereas the presence of the latter at the newly formed cross wall exceeds SCAMP5 following completion of cytokinesis. The results shown were observed in two independent lines. Scale bar indicates 5 µm. **j** Confocal analysis of SCAMP5-GFP vs ΔN-SCAMP5-GFP. An increased plasma membrane localization was observed in the absence of the double NPF motif. Colors are shown according to signal intensity (red to green to blue). The images are representative of two independent lines. Scale bar indicates 10 µm. **k** Quantification of the plasma membrane vs the cytoplasm of two independent lines for both constructs as shown in (**j**). The mean is shown as a pink dot. The amount of quantified cells is indicated below each boxplot. Different letters indicate significant differences between samples analyzed by one-sided Welch's ANOVA. Post hoc pairwise comparison was performed with the package multcomp utilizing the Tukey contrasts ($p \leq 0.001$). The boxplot extends from the 25th to 75th percentiles. The line inside the box marks the median. The whiskers go down and up to the 95% percentile.

Double NPF motifs or a single NPF motif flanked by an acidic stretch were identified before in yeast and mammalian systems and these motif extensions were shown to play a role in fine-tuning the interaction with the EH domains[15,16].

Additionally, we confirmed the interaction between AtEH1/Pan1 and SCAMP5 *in planta* using ratiometric bimolecular fluorescence complementation (rBiFC) in *N. benthamiana*. Confocal analysis confirmed the interaction between SCAMP5 and AtEH1/Pan1 specifically when both proteins were fused at their N-terminus with a half of YFP (Fig. 3g, h). The absence of interaction in case of C-terminal fusion proteins fits with the fact that both the EH1.1 domain in AtEH1/Pan1 as well as the NPF motif in SCAMP5 are located N-terminally. To challenge the rBiFC interaction, we repeated the experiment using AtEH1/Pan with a mutated EH1.1 binding pocket (W49A). The interaction was dramatically attenuated with fluorescence ratios that were only slightly higher than the ones observed for the C-terminally tagged proteins (Fig. 3g, h). Comparison of the NMR results on the NPF motif binding by the EH1.1 domain and the membrane-binding mode predicted by CG-MD simulations revealed that the binding sites are not mutually exclusive, i.e., EH1.1 can simultaneously interact with both proteins and lipids (Figs. 2f and 3e). Such a coincidence detection[40] could contribute to fine-tuning of membrane targeting of AtEH1/Pan1 and could further explain the observed differences in the localization of each domain. The nuclear envelope recruitment of the three-tandem EH1.1-GFP construct in *N. benthamiana* (Fig. 2c) could be an example of such a coincidence detection. Recently, PA was shown to reside at the nuclear envelope[41]. EH1.1 may, therefore, simultaneously interact with PA and other double NPF motif-containing proteins, such as the nuclear envelope resident protein NUP98A.

To address the physiological relevance of the interaction between SCAMP5 and AtEH1/Pan1 *in planta*, we imaged *A. thaliana* roots expressing SCAMP5-GFP and AtEH1/Pan1-mRuby3. SCAMP5 localizes mostly in endosomes and weakly at the plasma membrane (Supplementary Fig. 3). Co-localization with AtEH1/Pan1 at the PM was observed. Both proteins also prominently co-localize during various stages of cell plate formation where SCAMP5 clearly precedes the arrival of AtEH1/Pan1. However, the presence of AtEH1/Pan1 at the newly formed cross wall exceeds SCAMP5 following completion of cytokinesis (Fig. 3i). Altogether, our data suggest that SCAMP5 trafficking is highly dynamic. Short-term ES-9 treatment, a potent endocytic inhibitor[42], caused SCAMP5 accumulation at the PM (Supplementary Fig. 6), indicating the endocytic contribution to SCAMP5 dynamics. To further elucidate the role of the AtEH1/Pan1-SCAMP5 interaction in vivo, we compared the localization of the native SCAMP5 protein with an N-terminally truncated version (i.e. lacking the double NPF motif). In comparison to the wild type protein, the ΔN-SCAMP5 showed a reduced endosomal and an increased plasma membrane localization (Fig. 3j, k). These results corroborate our hypothesis that the double NPF motif is a recruitment signal that is involved in the retrograde transport of SCAMP5.

In conclusion, our parallel structural and functional comparison of the EH domains of AtEH/Pan1 revealed two divergent EH domains with differential, yet complementary functions. Both domains interact with negatively charged phospholipids, although they differ in their strength and specificity. In addition, the first EH domain binds a previously unidentified TPC interactor, SCAMP5, via its N-terminal double NPF motif. This constitutes a previously unidentified retrograde transport signal in plants. The tandem of EH domains in AtEH1/Pan1, a recurring leitmotif in the endocytic machinery in Eukarya, is a clear example of the evolutionary division of labor of a repetitive protein fold.

Methods

Multiple sequence alignment. To obtain protein sequences of AtEH/Pan1 homologs, GenBank (https://www.ncbi.nlm.nih.gov/genbank/), Joint Genome Institute (https://genome.jgi.doe.gov/portal/), EnsemblPlants (https://plants.ensembl.org/index.html) and Congenie (http://congenie.org/start) databases were used for a BLASTP search[43]. See Supplementary Data 5 for a complete list of all organisms searched (54 different plant genomes in total). A multiple alignment was constructed with the mafft algorithm in einsi mode[44]. Full sequence alignment can be found in Supplementary Data 5.

Protein production and purification. EH domains of AtEH1/Pan1 were amplified from the pDONR plasmid containing the full-length AtEH1/Pan1 coding sequence[3] and cloned by restriction digestion (Nde/Xho) into the pET22b plasmid. Primers used are shown in Supplementary Table 3. The final constructs have an N-terminal His-tag followed by a TEV-protease cleavage site and contain amino acids 1–107 (EH1.1) and 346–449 (EH1.2). To generate the tryptophan mutant in EH1.1, the complete plasmid was amplified using primers over the tryptophan-containing sequence. The fragment was re-assembled using NEBuilder® HiFi DNA Assembly Master Mix (NEB). To generate the EH1.2 R/K>E mutant, the construct was ordered as a gBlocks Gene Fragment (IDT) with (Nde/Xho) cleavage sites (Supplementary Table 4) and cloned into the pET22b plasmid as described above. Constructs were transformed into BL21(DE3) (#C2527H, NEB). Cells were grown at 37 °C in LB^+ medium and induced by the addition of 0.4 mM IPTG at OD 0.6 for 5 h. The yield of both domains was >5 mg/L culture. To obtain isotope-labeled proteins, cells were grown in M9 minimal medium supplemented with 0.5 g/l $N_{15}H_4Cl$ and/or 2 g/l ^{13}C-glucose (Eurisotop). Proteins were extracted using sonication in 20 mM HEPES pH 7.4, 150 mM NaCl, 2 mM $CaCl_2$, and Protease inhibitors (cOmplete ULTRA EDTA-free, Roche), except for proteins analyzed by the size exclusion chromatography multi-angle laser light scattering, for which no $CaCl_2$ was added during the purification. Purification was performed on an ÄKTA (GE Healthcare, Unicorn 3.10) system by purification using a HisPrep FF 16/10 (GE Healthcare). A one step elution was performed by adding 500 mM Imidazole during elution. This was followed by a gel filtration step using the same buffer as during the extraction but without the addition of protease inhibitors (HiLoad® 16/600 Superdex® 75 pg (GE Healthcare)). When no His-tag was required the protein was incubated overnight with 1/40 protein:his-TEV-protease (own production) at room temperature without shaking. Uncleaved protein and protease were removed via reverse IMAC (1 ml HisTRAP FF (GE Healthcare) followed by gel filtration (HiLoad® 16/600 Superdex® 75 pg (GE Healthcare)), both in the protein extraction buffer without the addition of protease inhibitors. The protein sequence of EH1.1 and EH1.2 along with their native molecular weight were verified by MS analysis.

GFP-His in an OPINF backbone (a generous gift from the lab of Ray Owen, OPPF, UK) was produced and purified as the EH domains without the addition of $CaCl_2$.

Multi-angle laser light scattering (MALLS). Purified His-tagged proteins EH1.1 (1 mg/ml) or EH1.2 (2 mg/ml) were injected onto a Superdex 75 Increase 10/300 GL size exclusion column (GE Healthcare), equilibrated with 20 mM HEPES pH 7.4, 300 mM NaCl, coupled to an online UV-detector (Shimadzu), a mini DAWN TREOS (Wyatt) multi-angle laser light scattering detector and an Optilab T-rEX refractometer (Wyatt) at room temperature. A refractive index increment (dn/dc) value of 0.185 ml/g was used. Band broadening corrections were applied using parameters derived from BSA injected under identical running conditions. Data analysis was carried out using the ASTRA6.1 software.

Protein crystallization of EH1.1. Commercial sparse matrix sitting drop crystallization screens were set up using a Mosquito liquid handling robot (TTP Labtech) using a 100 nl:100 nl, protein (12 mg/ml): mother liquor geometry in SwissSci 96-well triple drop plates. Plates were incubated at 293 K. An original hit in the JCSG screen (1.1 M SodiumMalonate, 0.1 M HEPES pH 7, 0.5% Jeffamine) was optimized to 100 mM HEPES pH 7.6, 0.8 M SodiumMalonate, 0.5% Jeffamine. Crystals were cryoprotected by the addition of ethylene glycol (15% v/v) to the mother liquor prior to plunging the crystals in liquid nitrogen for cryo-cooling prior to data collection.

Crystallographic structure determination. X-ray diffraction data were collected from single crystals at 100 K at the P14 microfocus beamline operated by the EMBL at PETRA III synchrotron (Hamburg, Germany). All data were integrated and scaled using XDS (Build20180126)[45]. The initial phases were generated by Automatic Molecular Replacement Pipeline (MoRDa)[46] using a search model derived from the X-ray structure of mouse EHD2 (2QPT) in combination with CCP4 (version 7.0). The initial structure was rebuilt with ARP/wARP[47] and further structure building and refinement was performed using Buster (version 2.10.3)[48] followed by iterative use of COOT[49] (version 0.8.8) and Phenix.refine[50] (version 1.12) software packages.

NMR structure determination of EH1.1 and EH1.2. For NMR structure determination the proteins were buffer exchanged using PD-10 columns (Sephadex G-25 M, GE Healthcare) or via gel filtration chromatography in the final purification step to 20 mM MES 6.5, 150 mM NaCl, 2 mM $CaCl_2$. All NMR spectra were recorded at CEITEC Josef Dadok National NMR Centre on 850 MHz Bruker Avance III spectrometer equipped with $^1H/^{13}C/^{15}N$ TCI cryogenic probe head with z-axis gradients. For each protein a set of three sparsely sampled 4D NMR experiments was acquired: 4D HC(CC-TOCSY(CO))NH, 4D $^{13}C,^{15}N$ edited HMQC-NOESY-HSQC (HCNH), and 4D $^{13}C,^{13}C$ edited HMQC-NOESY-HSQC (HCCH). Sequential and aliphatic side chain assignments were obtained automatically using the 4D-CHAINS algorithm that combines through-bond information from the 4D-TOCSY experiment and distance information from the 4D-NOESY (HCNH) experiment[23]. Aromatic sidechain frequencies were assigned manually by recording an additional 3D ^{13}C edited NOESY-HSQC experiment. Assignment completeness reached 99% for each EH domain. Backbone dihedral angle restraints were derived from TALOS using a combination of five kinds (H^N, H^α, C^α, C^β, N) of chemical shift assignments for each residue in the sequence[51]. NOE cross-peaks from the three NOESY spectra were assigned automatically by CYANA 3.0 in structure calculations with torsion angle dynamics[52]. Unambiguous distance restraints and torsion angle restraints (Supplementary Table 2) were used in a water refinement calculation[53] applying the RECOORD protocol[54]. The CNS patch introducing calcium coordination in a pentagonal bipyramidal configuration was prepared manually. CNS topology files for calcium coordination were generated based on the high-resolution crystal structure of calmodulin (PDB:1CLL). The quality of the NMR-derived structure ensembles was validated using PSVS[55] (version 1.4).

All-atom molecular dynamics simulations of Ca^{2+}-binding. All-atom molecular dynamics simulations were performed using the GROMACS 5.1.2 package[56]. We used the NMR structure of EH1.2 containing one calcium atom bound to the canonical motif (loop 2) as a starting structure for the simulation. We added a second calcium atom to the non-canonical motif (loop 1) asking whether it will be stably incorporated in the non-canonical motif and if so, what would be a stable coordination scheme. The same procedure was followed for EH1.2 Q382E. The simulation box contained one EH1.2 or EH1.2 Q382E molecule placed in a cubic box with a length of ~8 nm, which was filled with a 150 mM NaCl aqueous solution and which included additional Cl^- ions to neutralize the whole system. The protein and ions were parameterized using the CHARMM36 force field[57]. Water molecules were described with the TIP3P model[58]. Newton's equations of motion were integrated by employing the leap-frog algorithm with a time step of 2 fs. The trajectory frames were recorded every 10 ps. A cutoff of 1.2 nm was applied to short-range electrostatic interactions while long-range electrostatics was calculated with the use of the particle mesh Ewald method[59]. Van der Waals potentials were decreased so that the forces went smoothly to zero between 1.0 and 1.2 nm. Bonds with hydrogen atoms were constrained by the LINCS algorithm[60] and water molecules were kept rigid by the SETTLE algorithm[61]. The temperature of the system was maintained at 310 K using the velocity rescaling thermostat with a stochastic term[62] and the Parrinello–Rahman barostat[63] was utilized for semi-isotropic pressure coupling with a reference pressure of 1.01 bar. The time constants of the thermostat and barostat were 1 ps and 5 ps.

Coarse-grained molecular dynamics (CG-MD). The structure of EH1.1 or EH1.2 was mapped into the MARTINI CG representation using the martinize.py script[64]. The ELNEDYN representation with the distance cutoff 0.9 nm and the spring force constant 500 kJ mol^{-1} nm^{-2} was used to prevent any undesired large conformational changes during CG-MD simulations[65]. The MARTINI CG model for all lipid molecules used in this study was taken from Ingolfsson et al[66]. Lipid bilayer, in total composed of 440 phospholipid molecules, containing dioleoylphosphatidylcholine:dioleoylphosphatidylethanolamine:dioleoyl-phosphatidic acid (molecular ratio 2:2:1) was prepared using CharmmGUI Martini Maker[67]. CG-MD simulations were performed in GROMACS 5.1.2. The bond lengths were constrained to equilibrium lengths using the LINCS algorithm. Lennard-Jones and electrostatics interactions were cut off at 1.1 nm, with the potentials shifted to zero at the cutoff[64]. A relative dielectric constant of 15 was used. The neighbor list was updated every 20 steps using the Verlet neighbor search algorithm. Simulations were run in the NPT ensemble. The system was subject to pressure scaling to 1 bar using Parrinello–Rahman barostat with temperature scaling to 303 K using the velocity-rescaling method with coupling times of 1.0 and 12.0 ps. Simulations were performed using a 20 fs integration time step. Initially, the protein was placed 2.0 nm away from the membrane. Subsequently, the standard MARTINI water, Na^+ and Cl^- ions were added. The final concentration of NaCl was 150 mM. Additional Na^+ ions were added to the system to ensure the electroneutrality. The whole system was energy minimized using the steepest descent method up to the maximum of 500 steps, and equilibrated for 10 ns. Production runs were performed for up to 2 µs. The standard GROMACS tools, as well as in-house codes, were used for the analysis.

TXRF. The EH domains were buffer exchanged using PD-10 columns to a buffer containing 20 mM HEPES pH 7.4, 150 mM NaCl, 0.5 mM $CaCl_2$. The proteins were concentrated to 2 mM and the achieved protein concentration was verified by Nanodrop. TXRF quartz substrate disks were cleaned by placing them in a closed beaker with 5% HNO_3 solution under boiling conditions for half an hour. This cleaning process was then repeated using a 3% HNO_3 solution, rinsed twice using MilliQ H_2O and a final rinse using a MilliQ H_2O—ethanol solution and dried in vacuum. Five replicates for each EH domain consisting of 10 µl of the same protein solution were spotted on a quartz disk and dried under vacuum. The samples were measured in a G.N.R. TX2000 total reflection X-ray fluorescence spectrometer (40 kV, 30 mA, 1000 s LT, Mo anode). XRF data were fitted using the AXIL software package[68]. Ca–K_α integrated intensities were normalized for the Cl–K_α integrated intensities to account for small fluctuations in X-ray tube current and amount of probed sample volume.

Single-collector sector-field inductively coupled plasma-mass spectrometer (SF-ICP-MS). For each EH domain, the number of Ca^{2+} ions per molecule was determined via quantification of the calcium concentration in a solution containing a known number of protein molecules, as was determined by Nanodrop (Thermo Fisher). Calcium determination was accomplished using a Thermo Scientific Element XR (Bremen, Germany) SF-ICP-MS.

For this purpose, 10 µl volumes of EH1.1, of EH1.2 and of the buffer solution were diluted with 0.28 M HNO_3 (purified by sub-boiling distillation) and measured in triplicate. A buffer containing 150 mM of NaCl, 20 mM of HEPES at a pH of 7.4 and 0.5 mM of $CaCl_2$ was used as a procedure blank for the sample solutions containing 1.98 mM of EH1.1 and 1.94 mM of EH1.2, respectively. Sample preparation was carried out in a class-10 clean lab (PicoTrace™, Göttingen, Germany) at the A&MS-UGent research unit.

The samples were introduced into the ICP ion source using a 200 µl/min quartz nebulizer mounted onto a cyclonic spray chamber. Quantification was accomplished via external calibration. Standard solutions were prepared from 1000 mg/l calcium single-element standard solution (CPI International, Santa Rosa, CA, USA, lot 129901-31). The calibration curve for calcium was generated using five calibration standards, with concentrations ranging from 0 to 50 µg/l. Galium at a final concentration of 10 µg/l (originating from a 1000 mg/l single-element standard solution, Inorganic Ventures, Christiansburg, VA, USA) was used as an internal standard to correct for potential instrument instability and matrix effects. The signals for the nuclides ^{44}Ca and ^{69}Ga were measured at medium mass resolution (R = 4000) to avoid spectral overlap from CO_2^+ at a mass-to-charge ratio of 44. The instrument settings and data acquisition parameters used are summarized in Supplementary Data 1.

Liposome-binding experiments. For the liposome binding experiments, a vesicle co-sedimentation assay was used. Vesicle co-sedimentation assay were based on Kooijman et al.[69]. First, lipids were mixed in their specific ratios in chloroform, methanol was also added in the case of PI. Per sample, a total amount of 600 nmol lipids was used. Lipid mixtures were dried in a speedvac and stored at −20 °C. Second, Lipids were redissolved in 500 µl extrusion buffer (250 mM Raffinose pentahydrate, 1 mM DTT, 25 mM Tris-HCl pH 7.5), vortexed and incubated for 1 h after which the samples were sonicated for 1 min. Next, the lipid suspension was extruded twenty times over a polycarbonate membrane (pore size 0.2 µm). Before further use, the suspension was diluted in binding buffer (20 mM HEPES, pH 7.4, 150 mM NaCl and a variable amount of $CaCl_2$ was used) and spinned at 50,000 × *g* for 30 min at 22 °C. The liposome pellets were resuspended in 150 µl binding buffer. 0.5–1 µg of protein was diluted in 50 µl and added the the liposomes. The mix was incubated for 30–45 min at room temperature followed by 30 min centrifugation at 50,000 × *g*. Next, harvest the supernatant. Wash the pellet once by resuspending in 300–500 µl binding buffer, transferring the solution to a new vessel, and spin 20 min at 50,000 × *g*. Resuspend the pellet in 30 µl of the binding buffer and add 10 µl 6× SDS LB. To concentrate the supernatant 1 ml cold acetone (−20 °C) was added and incubated overnight. The next day, the supernatant pelleted by centrifugation at 20,000 × *g* for 10 min at 4 °C after which the pellets were redissolved in 30 µl of the binding buffer + 10 µl 6 x SDS LB. All samples were incubated at 95 °C for 5 min before loading on gel. Full gels are available in Supplementary Data 4.

Construction triple EH domains. Triple EH constructs were designed as a triple-tandem using the same amino acids sequence but variants in codon usage (*A. thaliana*, *Z. Mays*, *N. benthamiana*) separated by a triple GGGS-linker (Supplementary Table 5). At the start and end, 20 bp of pENTRY-L1-L2 plasmid[70] was added to clone the fragment into pENTRY-L1-L2 using the Gibson Assembly master mix (NEB). Next, the obtained plasmid was combined with pEN-L4-pH3.3-R1 and pDONRP2P3-EGFPst into a pFASTRK-II-m43GW backbone[71].

***N. Benthamiana* leaf infiltration**. *N. benthamiana* plants were grown in a greenhouse under long-day conditions (6–22 h light, 100 PAR, 21 °C) in soil (Saniflor osmocote pro NPK: 16-11-10 + magnesium and trace elements). Transient expression was performed by leaf infiltration according to Sparkes et al.[72]. A similar optical density of Agrobacterium strains was used for all constructs during co-expression. Transiently transformed *N. benthamiana* were imaged two to three days after infiltration Three-tandem EH domain constructs fused to GFP were expressed together with Cyclin D to induce cell division[27]. Imaging was performed on a PerkinElmer Ultraview spinning-disc system, attached to a Nikon Ti inverted microscope and operated using the Volocity software package (version 6.5.1). Images were acquired on an ImagEM ccd CCD camera (Hamamatsu C9100-13) using frame-sequential imaging with a ×60 water-immersion objective (NA = 1.20). Specific excitation and emission was performed using a 488 nm laser combined with a single band pass filter (500–550 nm) for GFP.

Construction of transgenic plants. SCAMP5 and ΔN-SCAMP5 cDNAs with LR gateway sites were generated using a BioXP printer (Supplementary Table 4) and cloned using LR clonase (Invitrogen) in pFASTRK-m43GW w/o terminator with a C-terminal GFP. Plant lines were generated by floral dip in Col-0 or complemented AtEH1/Pan1-mRuby3 Arabidopsis lines obtained from previous studies[3] (Supplementary Table 5). Primary transformants were selected by fluorescent selection of the seeds and the seedlings were checked for expression level on a confocal microscope. T2 lines were used for subsequent experiments.

Live cell imaging and chemical treatments. Root epidermal cells of 5–7 day old seedlings, grown vertically on 1/2 MS medium in continuous light were imaged on a Leica SP8X microscope (Leica, software 2019–2020) using the white-light laser with a 40x/1.1NA water-immersion lens.

For sequential dual-color imaging, EGFP was visualized using 488 nm laser excitation and a 495–550 nm spectral detection. mRuby3 was visualized using 558 nm laser excitation and a 600–700 nm spectral detection. Time gating was always applied except for imaging FM4-64.

For ES9 treatment plants were pretreated for 5 min with 10 µM ES9, a generous gift of the lab of Jenny Russinova (PSB, VIB/Ugent, BE), in 1/2 MS media followed by 30 min co-treatment of ES9 with 2 µM FM4-64.

rBiFC. rBiFC constructs were created in the 2in1 BiFC vectors[73]. rBIFC clones were amplified from the pDONR plasmid containing the full-length AtEH1/Pan1 coding sequence[3] or the mutated AtEH1/Pan1 coding sequence obtained through Gibson assembly (NEB) or obtained from previous rBIFC experiments[37]. Entry clones were combined in a Gateway LR recombination reaction with an empty BiFC destination vector and selected using LB containing spectinomycin and XgalI. Final BiFC vectors were checked by restriction digestion.

Ratiometric bimolecular fluorescence complementation (rBiFC) images were obtained using a Leica SP8X confocal microscope using the white-light laser with a 40×/1.1NA water-immersion objective. Images were acquired in line sequential mode, using 513 nm excitation and an emission window between 522 and 552 nm for YFP detection and 555 nm excitation and an emission window between 568 and 610 nm for RFP detection. All images were taken using the same ratio between YFP and RFP settings by adapting the strength of the white-light laser. The plasma membrane of the cells was selected by the active contour tool with a diameter of 5 px in ImageJ after which the average intensity of the RFP and YFP channels were selected. All images were devoid of saturated pixels.

PM/cytoplasm quantification. For the quantification of plasma membrane versus cytoplasm, the Fiji software package (v1.52p) was used. The 5% most intense pixels of the ROI, with a diameter of 5 px in ImageJ, covering the PM were divided by the 5% most intense pixels of the ROI inside of the cell. Only images devoid of saturated pixels were used for quantification.

Statistical analysis. For statistical analysis, the R and multcomp[74] package in R studio (version 1.4.1103) was used.

Peptidome profiling. Arabidopsis (Col-0) was grown vertically on 1/2 MS media for 7 days on a nylon mesh, harvested and flash-frozen in liquid nitrogen. Frozen seedlings (3 g/experiment) were ground using mortar and pestle. Proteins were extracted and denatured in 50 mM $NaHCO_3$, 8 M Urea and sonicated three times for 1 min. Protein extracts were rotated for 30 min at room temperature after which the extract was subsequently centrifuged twice at 20,000 × *g* for 20 min. DTT was added to the supernatant, was added to a final concentration of 5 mM and incubated at 55 °C for 30 min. Proteins were alkylated by the addition of 100 mM iodoacetamide and incubated for 15 min in the dark after which the mix was subsequently diluted with 50 mM $NaHCO_3$ to a final concentration of 2 M urea. Trypsin was added in a ratio of 1:75, protein:trypsin (Sequencing Grade Modified Trypsin, V5117, Promega) and incubated overnight at 37 °C on a rotating wheel.

Trypsin was removed using Sep-Pak Vac 3cc columns (500 mg, WAT036815, Waters). The extract was acidified using 1% TFA for 15 min on ice and cleared by centrifugation at 1780 × *g* for 15 min at room temperature. The cleared extract was split into four and applied on an equilibrated Sep-Pak columns (Waters, WAT036815, Lot # 010437235B). The column was pre-wet using 5 ml of 100% MeCN followed by sequential washing with 1 ml, 3 ml, and 6 ml of 0.1% TFA. After application of the extract, the column was washed sequentially with 1 ml, 5 ml, and 6 ml of 0.1% TFA followed by a 2 ml wash with 0.1% TFA, 5% acetonitrile. The peptides were eluted using three times 2 ml 0.1% TFA, 40% acetonitrile. The eluate was lyophilized for two days to remove TFA. 200 µg EH domain or his-GFP were coupled, in triplicate, during one hour to 25 µl Ni Sepharose 6 Fast Flow beads (GE Healthcare, 17-5318-01). Unbound protein was removed by three washing steps of 1 ml each. The lyophilized peptides were solubilized in 20 mM HEPES, 150 mM NaCl buffer, 0.2 mM $CaCl_2$, divided and added to the coupled beads. The peptidome was incubated with the proteins for 4 h after which the beads were washed three times with 1 ml of the binding buffer. Peptides were eluted by the addition of 80 µl 20 mM HEPES pH 8, 8 M Urea. The supernatant was removed from the beads and desalted with Monospin C18 columns (Agilent Technologies, A57003100) after which they were freeze dried[75].

Peptides were redissolved in 20 µl loading solvent A (0.1% TFA in water/acetonitrile (98:2, v/v)) of which 5 µl was injected for LC-MS/MS analysis on an Ultimate 3000 RSLC nano LC (Thermo Fisher Scientific, Bremen, Germany) in-line connected to a Q Exactive mass spectrometer (Thermo Fisher Scientific). The peptides were first loaded on a trapping column made in-house (100 µm internal

diameter (I.D.) × 20 mm, 5 µm beads C18 Reprosil-HD, Dr. Maisch, Ammerbuch-Entringen, Germany) and after flushing from the trapping column the peptides were separated on a 50 cm µPAC™ column with C18-endcapped functionality (Pharmafluidics, Belgium) kept at a constant temperature of 35 °C. Peptides were eluted by a linear gradient from 98% solvent A′ (0.1% formic acid in water) to 55% solvent B′ (0.1% formic acid in water/acetonitrile, 20/80 (v/v)) in 30 min at a flow rate of 300 nl/min, followed by a 5 min wash reaching 99% solvent B'. The mass spectrometer was operated in data-dependent, positive ionization mode, automatically switching between MS and MS/MS acquisition for the 5 most abundant peaks in a given MS spectrum. The source voltage was 2.2 kV, and the capillary temperature was 250 °C. One MS1 scan (*m/z* 400−2000, AGC target 3×10^6 ions, maximum ion injection time 80 ms), acquired at a resolution of 70,000 (at 200 *m/z*), was followed by up to 5 tandem MS scans (resolution 17,500 at 200 *m/z*) of the most intense ions fulfilling predefined selection criteria (AGC target 5×10^4 ions, maximum ion injection time 80 ms, isolation window 2 Da, fixed first mass 140 *m/z*, spectrum data type: centroid, intensity threshold $1.3 \times E4$, exclusion of unassigned, 1, 5–8, >8 positively charged precursors, peptide match preferred, exclude isotopes on, dynamic exclusion time 12 s). The HCD collision energy was set to 25% Normalized Collision Energy and the polydimethylcyclosiloxane background ion at 445.120025 Da was used for internal calibration (lock mass).

The raw files were processed with the MaxQuant software (version 1.6.4.0)[76], and searched with the built-in Andromeda search engine against the TAIR10_pep_20101214 database. Parameters can be found in Supplementary Data 2. Intensity values from the peptides output file of MaxQuant were used for quantitative analysis with the Perseus software (version 1.6.1.1). Intensity values were transformed to $\log^2$ values. Rows were filtered for at least two valid values in one of the sample groups, GFP, EH1.1, or EH1.2. Missing values were replaced with values from normal distribution with width of 0.3 and a downshift of 1.8. To determine the significantly enriched peptide sequences with the EH1 domains, a two-sided Student's *t*-test was performed between each of the EH domains versus GFP and the other EH domain as control. Permutation-based correction for multiple hypothesis testing was performed with thresholds FDR = 0.01 and S0 = 1.

NMR peptide binding. NPF peptide was dissolved in the final NMR buffer at a stock concentration of 4 mM. ^1H,^{15}N HSQC titrations of ^{15}N-labeled EH1.1 or EH1.2 with successive addition of unlabeled ligands were performed on samples containing 100 mM protein to a final concentration ratio of 1:4 excess of the peptide. Weighted chemical shift perturbations (CSP) were calculated as:

$$\mathrm{CSP} = \sqrt{\delta_{H^N}^2 + \frac{\delta_N^2}{6^2}} \quad (1)$$

The CSPs were fitted to a binding isotherm using the equation:

$$\mathrm{CSP} = \frac{\mathrm{CSP}_{\max}}{[2P_\mathrm{T}]}\left([L] + [P_\mathrm{T}] + K_\mathrm{D} - \sqrt{([L] + [P_\mathrm{T}] + K_\mathrm{D})^2 - 4[P_\mathrm{T}][L]}\right) \quad (2)$$

where CSP is the chemical shift perturbation at a given peptide concentration $[L]$, $\mathrm{CSP}_{\max}$ is the chemical shift perturbation at saturation, $[P_\mathrm{T}]$ is the total protein concentration, and K_D, the dissociation constant.

BLI. BLI experiments were performed on an Octet RED96 instrument (FortéBio, Octet Data acquisition software V12) using an HBS buffer supplemented with calcium (20 mM HEPES pH 7.4, 150 mM NaCl, 2 mM $CaCl_2$). A shake speed of 1000 rpm at 25 °C was used during all measurements. Ni-NTA (Molecular devices, 18-5102) biosensors were functionalized with EH domains (10 µg/ml) till a coupling signal of 3 nm was reached. The proteins were covalently coupled to the Ni-NTA biosensors by sequentially dipping in a 20 mM EDC:10 mM NHS mix for 60 s followed by 60 s quenching in 1 M ethanolamine pH 8.5. The coupled tips were equilibrated in the buffer before the addition of the analyte. Next, functionalized sensors were sequentially dipped in increasing concentrations of analyte with an association time of 60 s and a dissociation time of 120 s. To correct for bulk effects during the measurements we performed double reference subtraction. Here, non-functionalized sensors were exposed to the same analyte concentrations while a functionalized sensor was dipped in zero concentration of analyte. The reference traces were subtracted from the raw data before analysis. Req values used in the analysis were determined for each concentration by averaging 30 data points once the sensors achieved stable equilibrium. Graphpad Prism 5 was used to analyse and plot binding data using the one-site total binding model. Peptides were ordered from Peptide 2.0 (95% purity).

Visualisation of protein structures and data. Bio-Rad Image Lab imager (version 6.0.0 Build 25) was used to scan SDS-PAGE gels. For the visualisation of all protein structures UCSF Chimera (version 1.13.1) was used. Secondary structure of protein models was assigned using the DSSP web server (version 3.0). Mapping of conserved residues was performed by combining the Consurf server[77] with the generated alignment for the individual domains. All figures were prepared utilizing the Inkscape program (https://inkscape.org/).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets generated during and/or analyzed during the current study are available as Source data, Supplementary Data files or have been deposited to the Protein Data Bank (pdbe.org). Other data are available from the corresponding author upon reasonable request. The structural data that support the findings of this study are available in the Protein Data Bank with the accession codes 6YIG, 6YEU and 6YET. The following structures were used 2QPT, 1CLL, 2KSP, 1F8H. Source data are provided with this paper.

Received: 19 May 2020; Accepted: 22 April 2021;
Published online: 24 May 2021

References

1. Kaksonen, M. & Roux, A. Mechanisms of clathrin-mediated endocytosis. *Nat. Rev. Mol. Cell Biol.* **19**, 313–326 (2018).
2. Qi, X., Pleskot, R., Irani, N. G. & van Damme, D. Meeting report—cellular gateways: expanding the role of endocytosis in plant development. *J. Cell Sci.* **131**, jcs222604 (2018).
3. Gadeyne, A. et al. The TPLATE adaptor complex drives clathrin-mediated endocytosis in plants. *Cell* **156**, 691–704 (2014).
4. di Rubbo, S. et al. The clathrin adaptor complex AP-2 mediates endocytosis of brassinosteroid insensitive1 in Arabidopsis. *Plant Cell* **25**, 2986–2997 (2013).
5. Hirst, J. et al. Characterization of TSET, an ancient and widespread membrane trafficking complex. *eLife* **3**, e02866 (2014).
6. Yperman, K. et al. Molecular architecture of the endocytic TPLATE complex. *Sci. Adv.* **7**, eabe7999 (2021).
7. Wang, J. et al. High temporal resolution reveals simultaneous plasma membrane recruitment of the TPLATE complex subunits. *Plant Physiol.* **183**, 986–997 (2020).
8. Narasimhan, M. et al. Evolutionarily unique mechanistic framework of clathrin-mediated endocytosis in plants. *eLife* **9**, e52067 (2020).
9. Duncan, M. C., Cope, M. J. T. V., Goode, B. L., Wendland, B. & Drubin, D. G. Yeast Eps15-like endocytic protein, Pan1p, activates the Arp2/3 complex. *Nat. Cell Biol.* **3**, 687–690 (2001).
10. Toshima, J., Toshima, J. Y., Martin, A. C. & Drubin, D. G. Phosphoregulation of Arp2/3-dependent actin assembly during receptor-mediated endocytosis. *Nat. Cell Biol.* **7**, 246–254 (2005).
11. Toshima, J. et al. Negative regulation of yeast Eps15-like Arp2/3 complex activator, Pan1p, by the Hip1R-related protein, Sla2p, during endocytosis. *Mol. Biol. Cell* **18**, 658–668 (2007).
12. Toshima, J. Y. et al. Yeast Eps15-like endocytic protein Pan1p regulates the interaction between endocytic vesicles, endosomes and the actin cytoskeleton. *Elife* **5**, e10276 (2016).
13. Wang, P. et al. Plant AtEH/Pan1 proteins drive autophagosome formation at ER-PM contact sites with actin and endocytic machinery. *Nat. Commun.* **10**, 5132 (2019).
14. de Beer, T. et al. Molecular mechanism of NPF recognition by EH domains. *Nat. Struct. Mol. Biol.* **7**, nsb1100_1018 (2000).
15. Kieken, F. et al. Structural insight into the interaction of proteins containing NPF, DPF, and GPF motifs with the C-terminal EH-domain of EHD1. *Protein Sci.* **18**, 2471–2479 (2009).
16. Henry, G. D., Corrigan, D. J., Dineen, J. V. & Baleja, J. D. Charge effects in the selection of NPF motifs by the EH domain of EHD1. *Biochemistry* **49**, 3381–3392 (2010).
17. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* **46**, gkx922– (2017).
18. Sigrist, C. J. A. et al. New and continuing developments at PROSITE. *Nucleic Acids Res.* **41**, D344–D347 (2013).
19. Bar, M. et al. AtEHDs, novel Arabidopsis EH-domain-containing proteins involved in endocytosis. *Plant J.* **55**, 1025–1038 (2008).
20. Bar, M. & Avni, A. EHD2 inhibits signaling of Leucine rich repeat receptor-like proteins. *Plant Signal. Behav.* **4**, 682–684 (2009).
21. Bar, M., Leibman, M., Schuster, S., Pitzhadza, H. & Avni, A. EHD1 Functions in endosomal recycling and confers salt tolerance. *PLoS ONE* **8**, e54533 (2013).
22. Sánchez-Rodríguez, C. et al. The cellulose synthases are cargo of the TPLATE adaptor complex. *Mol. Plant* **11**, 346–349 (2018).
23. Evangelidis, T. et al. Automated NMR resonance assignments and structure determination using a minimal set of 4D spectra. *Nat. Commun.* **9**, 384 (2018).
24. Naslavsky, N., Rahajeng, J., Chenavas, S., Sorgen, P. L. & Caplan, S. EHD1 and Eps15 Interact with phosphatidylinositols via their Eps15 homology domains. *J. Biol. Chem.* **282**, 16612–16622 (2007).

25. Paoluzi, S. et al. Recognition specificity of individual EH domains of mammals and yeast. *EMBO J.* **17**, 6541–6550 (1998).
26. Bilkova, E. et al. Calcium directly regulates phosphatidylinositol 4,5-bisphosphate headgroup conformation and recognition. *J. Am. Chem. Soc.* **139**, 4019–4024 (2017).
27. Xu, J., Lee, Y. R. J. & Liu, B. Establishment of a mitotic model system by transient expression of the D-type cyclin in differentiated leaf cells of tobacco (Nicotiana benthamiana). *N. Phytologist* **226**, 1213–1220 (2020).
28. Simon, M. L. A. et al. A PtdIns(4)P-driven electrostatic field controls cell membrane identity and signalling in plants. *Nat. Plants* **2**, 16089 (2016).
29. Platre, M. P. et al. A combinatorial lipid code shapes the electrostatic landscape of plant endomembranes. *Dev. Cell* **45**, 465–480.e11 (2018).
30. McLoughlin, F. et al. Identification of novel candidate phosphatidic acid-binding proteins involved in the salt-stress response of Arabidopsis thaliana roots. *Biochemical J.* **450**, 573–581 (2013).
31. Kaneda, M. et al. Plant AP180 N-terminal homolog proteins are involved in clathrin-dependent endocytosis during pollen tube growth in arabidopsis thaliana. *Plant Cell Physiol.* **60**, 1316–1330 (2019).
32. Putta, P., Creque, E., Piontkivska, H. & Kooijman, E. E. Lipid−protein interactions for ECA1 an N-ANTH domain protein involved in stress signaling in plants. *Chem. Phys. Lipids* **231**, 104919 (2020).
33. Fernandez-Chacon, R., Sudhof, T. C., Ferna, R. & Su, T. C. Novel SCAMPs lacking NPF repeats: Ubiquitous and synaptic vesicle-specific forms implicate SCAMPs in multiple membrane-trafficking functions. *J. Neurosci.* **20**, 7941–7950 (2000).
34. Bourdais, G. et al. The use of quantitative imaging to investigate regulators of membrane trafficking in Arabidopsis stomatal closure. *Traffic* **20**, 168–180 (2019).
35. Sheung, K. L., Cai, Y., Hillmer, S., Robinson, D. G. & Jiang, L. SCAMPs highlight the developing cell plate during cytokinesis in tobacco BY-2 cells. *Plant Physiol.* **147**, 1637–1645 (2008).
36. Park, D. et al. Impairment of release site clearance within the active zone by reduced SCAMP5 expression causes short-term depression of synaptic release. *Cell Rep.* **22**, 3339–3350 (2018).
37. Arora, D. et al. Establishment of proximity-dependent biotinylation approaches in different plant model systems. *Plant Cell* 32, 3388–3407 (2020).
38. Bateman, A. et al. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, gky1049– (2018).
39. Rumpf, J. et al. Structure of the Eps15-stonin2 complex provides a molecular explanation for EH-domain ligand specificity. *EMBO J.* **27**, 558–569 (2008).
40. Moravcevic, K., Oxley, C. L. & Lemmon, M. A. Conditional peripheral membrane proteins: facing up to limited specificity. *Structure* **20**, 15–27 (2012).
41. Thaller, D. J. et al. Direct binding of ESCRT protein Chm7 to phosphatidic acid–rich membranes at nuclear envelope herniations. *J. Cell Biol.* **220**, e202004222 (2021).
42. Dejonghe, W. et al. Mitochondrial uncouplers inhibit clathrin-mediated endocytosis largely through cytoplasmic acidification. *Nat. Commun.* **7**, 11710 (2016).
43. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
44. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinforma.* **20**, 1160–1166 (2017).
45. Kabsch, W. et al. XDS. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **66**, 125–132 (2010).
46. Vagin, A. & Lebedev, A. MoRDa, an automatic molecular replacement pipeline. *Acta Crystallogr. Sect. A Found. Adv.* **71**, s19–s19 (2015).
47. Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat. Protoc.* **3**, 1171–1179 (2008).
48. Bricogne, G. et al. *BUSTER version 2.10.3*. (Global Phasing Ltd., 2011).
49. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. Sect. D: Biol. Crystallogr.* **66**, 486–501 (2010).
50. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D: Biol. Crystallogr.* **66**, 213–221 (2010).
51. Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. TALOS+: A hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomolecular NMR* **44**, 213–223 (2009).
52. Güntert, P. Automated structure determination from NMR spectra. *Eur. Biophys. J.* **38**, 129–143 (2009).
53. Linge, J. P., Williams, M. A., Spronk, C. A. E. M., Bonvin, A. M. J. J. & Nilges, M. Refinement of protein structures in explicit solvent. *Proteins: Struct., Funct., Bioinforma.* **50**, 496–506 (2003).
54. Nederveen, A. J. et al. RECOORD: A recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. *Proteins: Struct., Funct. Genet.* **59**, 662–672 (2005).
55. Bhattacharya, A., Tejero, R. & Montelione, G. T. Evaluating protein structures determined by structural genomics consortia. *Proteins: Struct., Funct., Bioinforma.* **66**, 778–795 (2006).
56. Abraham, M. J. et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
57. Huang, J. & MacKerell, A. D. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J. Comput. Chem.* **34**, 2135–2145 (2013).
58. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
59. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N ·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
60. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).
61. Miyamoto, S. & Kollman, P. A. Settle: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **13**, 952–962 (1992).
62. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 14101 (2007).
63. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
64. de Jong, D. H. et al. Improved parameters for the martini coarse-grained protein force field. *J. Chem. Theory Comput.* **9**, 687–697 (2013).
65. Periole, X., Cavalli, M., Marrink, S.-J. & Ceruso, M. A. Combining an elastic network with a coarse-grained molecular force field: structure, dynamics, and intermolecular recognition. *J. Chem. Theory Comput.* **5**, 2531–2543 (2009).
66. Ingólfsson, H. I. et al. Lipid organization of the plasma membrane. *J. Am. Chem. Soc.* **136**, 14554–14559 (2014).
67. Hsu, P.-C. et al. CHARMM-GUI Martini Maker for modeling and simulation of complex bacterial membranes with lipopolysaccharides. *J. Comput. Chem.* **38**, 2354–2363 (2017).
68. Vekemans, B., Janssens, K., Vincze, L., Adams, F. & van Espen, P. Analysis of X-ray spectra by iterative least squares (AXIL): New developments. *X-Ray Spectrom.* **23**, 278–285 (1994).
69. Kooijman, E. E. et al. An electrostatic/hydrogen bond switch as the basis for the specific interaction of phosphatidic acid with proteins. *J. Biol. Chem.* **282**, 11356–11364 (2007).
70. Houbaert, A. et al. POLAR-guided signalling complex assembly and localization drive asymmetric cell division. *Nature* **563**, 574–578 (2018).
71. Karimi, M. & Jacobs, T. B. GoldenGateway: a DNA assembly method for plant biotechnology. *Trends Plant Sci.* **26**, 95–96 (2021).
72. Sparkes, I. A., Runions, J., Kearns, A. & Hawes, C. Rapid, transient expression of fluorescent fusion proteins in tobacco plants and generation of stably transformed plants. *Nat. Protoc.* **1**, 2019–2025 (2006).
73. Grefen, C. & Blatt, M. R. A 2in1 cloning system enables ratiometric bimolecular fluorescence complementation (rBiFC). *BioTechniques* **53**, 311–314 (2012).
74. Herberich, E., Sikorski, J. & Hothorn, T. A robust procedure for comparing multiple means under heteroscedasticity in unbalanced designs. *PLoS ONE* **5**, e9788 (2010).
75. van Leene, J. et al. Capturing the phosphorylation and protein interaction landscape of the plant TOR kinase. *Nat. Plants* **5**, 316–327 (2019).
76. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
77. Ashkenazy, H. et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**, W344–W350 (2016).

Acknowledgements

We would like to thank Prof. Eugenia Russinova (Ghent University/VIB-PSB) for an aliquot of the ES9 compound, Prof. Ray Owens (OPPF, Research Complex, Harwell) for initial work on the EH constructs (PID:1724/1846) and an aliquot of the GFP-His plasmid. We also express our gratitude to the proteomics core facility of VIB for the help and expertise with running all MS experiments. This research in the D.V.D. lab is supported by the European Research Council T-Rex project number 682436 and by the National Science Foundation Flanders (FWO; G009415N, 3G017919). The research in the K.T. lab is supported by project CEITEC 2020 (no. LQ1601) with financial contribution from the MEYS CR and National Programme for Sustainability II and by Grant Agency of Masaryk University (MUNI/G/0739/2017). CIISB research infrastructure project LM2018127 funded by MEYS CR is gratefully acknowledged for the financial support of the measurements at CEITEC Josef Dadok National NMR Centre. We thank the staff of beamlines P14 (PETRAIII) and Proxima2A (SOLEIL) for beam time allocation and excellent technical support. Y.B. was supported by a post-doctoral research fellowship from Research Foundation Flanders (FWO, 3E002518, Belgium). R.P. and M.

P. are supported by the Czech Science Foundation grant 19-21758S. S.D.M. was supported by a pre-doctoral fellowship from the Flanders Agency for Innovation and Entrepreneurship (VLAIO-Flanders, Belgium). S.N.S. acknowledges research support from the Hercules Foundation (no. AUGE- 11-029), Ghent University (BOF17-GOA-028) and VIB. Q.J. is supported by the China Scholarship Council (201906760018). Panel a of Fig. 3 was created with BioRender.com. We would like to thank iNEXT (PID: 6554) for funding the structure determination of EH1.1.

Author contributions

K.Y. and R.P. designed and performed most of the experiments and wrote the paper with S.N.S., K.T., and D.V.D. A.P. performed NMR structure calculations calcium precipitation experiments (NMR) and peptide titration experiments (NMR); T.E. helped with structure calculations; R.M. performed MALLS and X-ray structure determination and helped in designing EH constructs and protein crystallization; S.D.M. performed BLI; Y.B. collected X-ray data and assisted in structure determination; D.E. performed MS analysis; Q.J. helped with characterization of the SCAMP5 lines and rBiFC; M.P. helped in performing liposome binding assays; P.T. performed TXRF; R.G. performed the ICP-MS experiment. K.T., S.N.S., P.V., G.D.J., F.V., L.V., J.V.L., R.P. and D.V.D. were responsible for experimental design, research supervision and finalizing the manuscript text.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-23314-6.

Correspondence and requests for materials should be addressed to S.N.S., K.T., R.P. or D.V.D.

Peer review information *Nature Communications* thanks Richard Baker and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021

nature structural & molecular biology

Structural basis for dimethylarginine recognition by the Tudor domains of human SMN and SPF30 proteins

Konstantinos Tripsianes[1,2], Tobias Madl[1,2], Martin Machyna[3], Dimitrios Fessas[4], Clemens Englbrecht[5], Utz Fischer[5], Karla M Neugebauer[3] & Michael Sattler[1,2]

Arginine dimethylation plays critical roles in the assembly of ribonucleoprotein complexes in pre-mRNA splicing and piRNA pathways. We report solution structures of SMN and SPF30 Tudor domains bound to symmetric and asymmetric dimethylated arginine (DMA) that is inherent in the RNP complexes. An aromatic cage in the Tudor domain mediates dimethylarginine recognition by electrostatic stabilization through cation-π interactions. Distinct from extended Tudor domains, dimethylarginine binding by the SMN and SPF30 Tudor domains is independent of proximal residues in the ligand. Yet, enhanced micromolar affinities are obtained by external cooperativity when multiple methylation marks are presented in arginine- and glycine-rich peptide ligands. A hydrogen bond network in the SMN Tudor domain, including Glu134 and a tyrosine hydroxyl of the aromatic cage, enhances cation-π interactions and is impaired by a mutation causing an E134K substitution associated with spinal muscular atrophy. Our structural analysis enables the design of an optimized binding pocket and the prediction of DMA binding properties of Tudor domains.

© 2011 Nature America, Inc. All rights reserved.

Arginine methylation has long been observed in RNA-binding proteins and histones, but it has only recently been implicated in a variety of cellular processes, including RNA processing, transcriptional regulation, trafficking, signal transduction and gametogenesis[1,2]. Members of a broad family of protein arginine methyltransferases (PRMTs) generate asymmetric DMA (aDMA; PRMT1) by placing two methyl groups on one of the terminal guanidino nitrogen atoms of arginine, or symmetric DMA (sDMA; PRMT5) by placing one methyl group on each of the terminal guanidino nitrogen atoms, whereas monomethylated arginine (MMA) is probably generated as an intermediate *en route* to dimethylation[1]. Despite the increasing number of protein targets for arginine methylation[3], the only modules known so far to interpret dimethylarginine signatures are the Tudor domains of survival motor neuron protein (SMN), survival of motor neuron–related splicing factor 30 (SPF30), staphylococcal nuclease domain–containing protein 1 (SND1; also known as p100) and Tudor domain–containing protein (TDRD) family members[4–7]. SMN and SPF30 (**Fig. 1a**) play critical roles in the assembly of uridine-rich small nuclear ribonucleoprotein complexes (U snRNPs)[8–10] and in pre-mRNA splicing[11,12], respectively. U snRNPs have in common a seven-membered ring of Sm proteins, which forms around a conserved RNA sequence. Both SMN and SPF30 contain single Tudor domains that have been shown to associate directly with a subset of Sm proteins through sDMA modifications in their arginine- and glycine-rich C-terminal tails (**Fig. 1b** and refs. 4–6,13–16). In addition, the presence of SMN in Cajal bodies (spherical suborganelles found in the nuclei of proliferative or metabolically active cells) has been shown to depend on sDMA modifications of an arginine- and glycine-rich region located at the C-terminal end of coilin[17] (one of the main components of Cajal bodies), suggesting that recognition of sDMA marks in coilin determines SMN's subnuclear localization in Cajal bodies[18].

More recently, it was found that PIWI-ribonucleoprotein complexes (piRNPs) consisting of PIWI proteins and PIWI-interacting RNAs (piRNAs) can silence transposable elements in the germline during early development. Biogenesis of piRNAs depends on the recognition of sDMA modifications in the N-terminal regions of PIWI proteins (**Fig. 1b**) by TDRD Tudor domains[2,19–21]. A common feature of Tudor proteins that participate in the piRNA pathway is an extended architecture, whereby the basic Tudor domain is fused to an oligonucleotide- and oligosaccharide-binding (OB) fold. Two recent structural studies have underscored the requirement of the OB fold in assisting the Tudor domain for binding sDMA-containing peptides by interaction with nonmethylated residues adjacent to the modification[22,23].

Here, we set out to address the structural and energetic determinants of dimethylarginine recognition by the prototypic Tudor domains of SMN and SPF30. These proteins contain only the basic Tudor fold, which alone is sufficient for binding arginine- and glycine-rich peptides containing sDMA and aDMA with low micromolar affinity. Our study highlights unprecedented features in the recognition mode of arginine dimethylation and should greatly assist in understanding the evolution and functional role of the Tudor fold.

[1]Institute of Structural Biology, Helmholtz Zentrum München, Neuherberg, Germany. [2]Munich Center for Integrated Protein Science at the Chair of Biomolecular NMR, Department Chemie, Technische Universität München, Garching, Germany. [3]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany. [4]Dipartimento di Tecnologie Alimentari e Microbiologiche, Università di Milano, Milano, Italy. [5]Department of Biochemistry, Theodor Boveri Institute, University of Würzburg, Würzburg, Germany. Correspondence should be addressed to M.S. (sattler@helmholtz-muenchen.de).

Received 25 August 2010; accepted 14 October 2011; published online 20 November 2011; doi:10.1038/nsmb.2185

RESULTS

Thermodynamic analysis of Tudor-dimethylarginine interactions

Binding assays using isothermal titration calorimetry (ITC) demonstrated that the SMN and SPF30 Tudor domains do not bind arginine, as well as lysine or its (mono-, di- and tri-) methylated forms (data not shown). Arginine methylation, on the other hand, induced measurable affinities, with binding increasing from MMA to aDMA to sDMA (**Fig. 1** and **Supplementary Table 1a**). SMN showed higher affinity for all ligands compared with SPF30. Binding affinities for an Sm B/B′-derived pentapeptide harboring a single sDMA residue were nearly equivalent to those of the isolated sDMA. Similar binding affinities and stoichiometries were also observed for the Sm D1 hexapeptide containing two sDMAs. Thus, a second sDMA motif separated by a single glycine residue did not enhance the binding affinity (**Supplementary Fig. 1**). However, the picture changed substantially when the long peptide derived from the Sm D3 tail, which harbors four sDMA

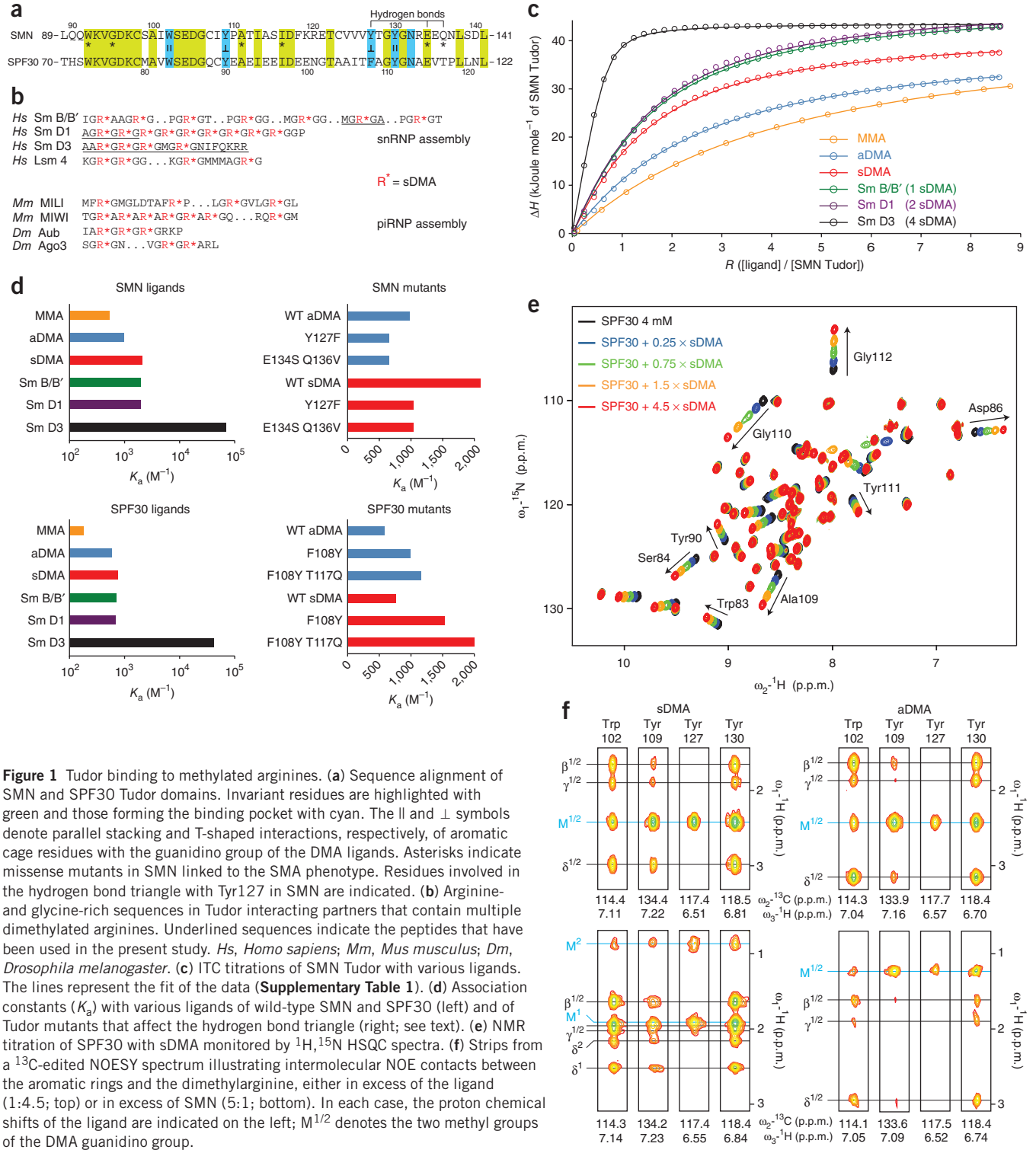


Figure 1 Tudor binding to methylated arginines. (**a**) Sequence alignment of SMN and SPF30 Tudor domains. Invariant residues are highlighted with green and those forming the binding pocket with cyan. The ∥ and ⊥ symbols denote parallel stacking and T-shaped interactions, respectively, of aromatic cage residues with the guanidino group of the DMA ligands. Asterisks indicate missense mutants in SMN linked to the SMA phenotype. Residues involved in the hydrogen bond triangle with Tyr127 in SMN are indicated. (**b**) Arginine- and glycine-rich sequences in Tudor interacting partners that contain multiple dimethylated arginines. Underlined sequences indicate the peptides that have been used in the present study. *Hs*, *Homo sapiens*; *Mm*, *Mus musculus*; *Dm*, *Drosophila melanogaster*. (**c**) ITC titrations of SMN Tudor with various ligands. The lines represent the fit of the data (**Supplementary Table 1**). (**d**) Association constants (K_a) with various ligands of wild-type SMN and SPF30 (left) and of Tudor mutants that affect the hydrogen bond triangle (right; see text). (**e**) NMR titration of SPF30 with sDMA monitored by $^{1}H,^{15}N$ HSQC spectra. (**f**) Strips from a ^{13}C-edited NOESY spectrum illustrating intermolecular NOE contacts between the aromatic rings and the dimethylarginine, either in excess of the ligand (1:4.5; top) or in excess of SMN (5:1; bottom). In each case, the proton chemical shifts of the ligand are indicated on the left; $M^{1/2}$ denotes the two methyl groups of the DMA guanidino group.

© 2011 Nature America, Inc. All rights reserved.

Table 1 NMR and refinement statistics for the Tudor–DMA complexes

	SMN–sDMA	SPF30–sDMA	SMN–aDMA	SPF30–aDMA
NMR distance and dihedral restraints				
Distance restraints				
Total NOE	1,865	1,747	1,862	1,811
Intra-residue	374	310	382	332
Inter-residue	1,491	1,437	1,480	1,479
Sequential ($\|i-j\|=1$)	446	456	440	464
Medium-range ($\|i-j\|<4$)	246	256	239	271
Long-range ($\|i-j\|>5$)	724	646	726	671
Intermolecular	75	79	75	73
Hydrogen bonds	3		3	
Total dihedral angle restraints	85	97	85	97
ϕ	42	47	42	47
ψ	43	50	43	50
Structure statistics				
Violations (mean ± s.d.)				
Distance restraints (Å)	0.010 ± 0.00	0.009 ± 0.00	0.010 ± 0.00	0.009 ± 0.00
Dihedral angle restraints (°)	0.545 ± 0.05	0.542 ± 0.07	0.754 ± 0.06	0.521 ± 0.06
Max. dihedral angle violation (°)	2.68	3.05	4.73	3.05
Max. distance restraint violation (Å)	0.27	0.39	0.18	0.18
Deviations from idealized geometry				
Bond lengths (Å)	0.013 ± 0.00	0.012 ± 0.00	0.013 ± 0.00	0.013 ± 0.00
Bond angles (°)	1.206 ± 0.03	1.058 ± 0.03	1.156 ± 0.02	1.023 ± 0.03
Impropers (°)	1.204 ± 0.06	1.280 ± 0.05	1.211 ± 0.05	1.197 ± 0.08
Average pairwise r.m.s. deviation (Å)[a]				
Heavy	0.67 ± 0.10	0.67 ± 0.10	0.65 ± 0.09	0.67 ± 0.09
Backbone	0.18 ± 0.04	0.24 ± 0.05	0.19 ± 0.05	0.23 ± 0.05

[a]Pairwise r.m.s. deviation was calculated among 20 refined structures for residues 92–141 of SMN and 73–122 of SPF30.

© 2011 Nature America, Inc. All rights reserved.

marks distributed over a range of ten residues (**Fig. 1b**), was studied. Both SMN and SPF30 Tudor domains bound to this ligand with 50× the strength of the shorter peptides containing only one or two sDMA modifications (Sm B/B′, Sm D1), with affinities in the low micromolar range (**Fig. 1c**,**d** and **Supplementary Table 1a**). These affinities were comparable to the ones reported for other Tudor domains (tandem, hybrid or extended forms) that bind methylation marks (lysine or arginine) but require contacts with proximal residues for efficient recognition[22–26]. A series of NMR titrations using the SMN Tudor domain (residues 84–147) and a longer construct (residues 3–160) demonstrated that neither the unstructured region around the SMN Tudor fold nor the residues flanking the methylation marks contributed to binding (**Supplementary Fig. 1** and ref. 16). Analysis of the ITC data indicated that binding enhancement to the long peptide was of entropic nature (**Supplementary Fig. 2**). As the aromatic cage of the Tudor domain can only accommodate a single DMA residue (see below), the strongly increased binding likely reflected external cooperativity effects induced by the increased local concentration of multiple sDMAs presented in the long peptide (**Fig. 1c**,**d**). Consistently, known SMN-Tudor ligands comprise several sDMA modifications in arginine- and glycine-rich peptide motifs[4], implying that this is required for high (micromolar K_d) binding affinity.

Tudor NMR structures in complex with sDMA and aDMA

In NMR titrations, the SMN and SPF30 Tudor domains bound sDMA and aDMA in the fast-exchange regime on the NMR chemical shift time scale (**Fig. 1e** and **Supplementary Fig. 3**), consistent with the binding affinities measured by ITC (**Supplementary Table 1a**). Based on a large number of total and intermolecular distance restraints for the SMN–sDMA, SPF30–sDMA, SMN–aDMA and SPF30–aDMA complexes (**Fig. 1f**, **Table 1** and **Supplementary Fig. 4**), we calculated high-quality structures (**Fig. 2**, **Table 1** and **Supplementary Fig. 5**). The characteristic strongly curved β-sheet fold of the Tudor domain did not show any appreciable structural perturbation upon binding to sDMA and aDMA (**Supplementary Fig. 5**). In either complex, the planar guanidino group of the dimethylated arginine was sandwiched between the aromatic side chains of a tryptophan and a tyrosine residue, whereas another tyrosine and a fourth aromatic residue (SMN Tyr127 and SPF30 Phe108) were positioned orthogonally to this arrangement constituting the aromatic cage (**Figs. 1a** and **2b**). An asparagine residue provided additional van der Waals contacts to the ligand and completed the binding pocket (**Fig. 2b**). Both sDMA and aDMA inserted into the cavity with their methyl groups facing the aromatic side chain (tyrosine or phenylalanine) located at the far end, as confirmed by the NOE contacts (**Supplementary Fig. 6**). Arginine dimethylation imparted hydrophobicity and bulkiness to the residue, but it did not neutralize the cationic charge (**Fig. 2a** and **Supplementary Table 2**). The increased bulk and hydrophobicity facilitated nonelectrostatic contacts with the cage and, most importantly, placed the cationic carbon in a favorable position for cation-π interactions (**Fig. 2c**). For sDMA, the cation was stacked parallel at van der Waals distance from the tryptophan and tyrosine ring centroids (face of the aromat with face of the guanidino group), maximizing cation-π stabilization. In addition, it occupied the focal point of the π-electron system of the two other aromatic side chains that line the cavity (**Fig. 2b**,**c**), thereby contributing additional cation-π interactions in a T-shaped geometry (face of the aromat with edge of the guanidino group)[27,28]. Such a coordination scheme efficiently utilized the polar interactions in an otherwise hydrophobic environment (**Fig. 2b**,**c**). The aDMA cation, on the other hand, was not perfectly perpendicular to the aromatic rings in the stacking geometry, thus reducing the electrostatic stabilization compared with sDMA, an observation consistent with the lower binding affinity. The T-shaped interactions with the other two aromatic side chains became more relevant due to shorter distances. As a common principle of DMA recognition, the steric bulk provided by the addition of two methyl groups to the arginine side chain guides the position of the cation relative to the rigid aromatic cage. This mode of recognition achieves a specific readout of the dimethylation marks and discriminates against nonmethylated and monomethylated arginines.

Contributions of aromatic cage residues to DMA specificity

To assess the contributions of the aromatic cage residues toward methylarginine binding, we individually replaced these in SMN Tudor with amino acids that are frequently observed in the sequences of human TDRD and the *Drosophila* TUDOR proteins (**Fig. 3a** and **Supplementary Fig. 7**). Substitution of any of the four aromatic residues by polar or aliphatic side chains abrogated binding. N132D had mild effects, and

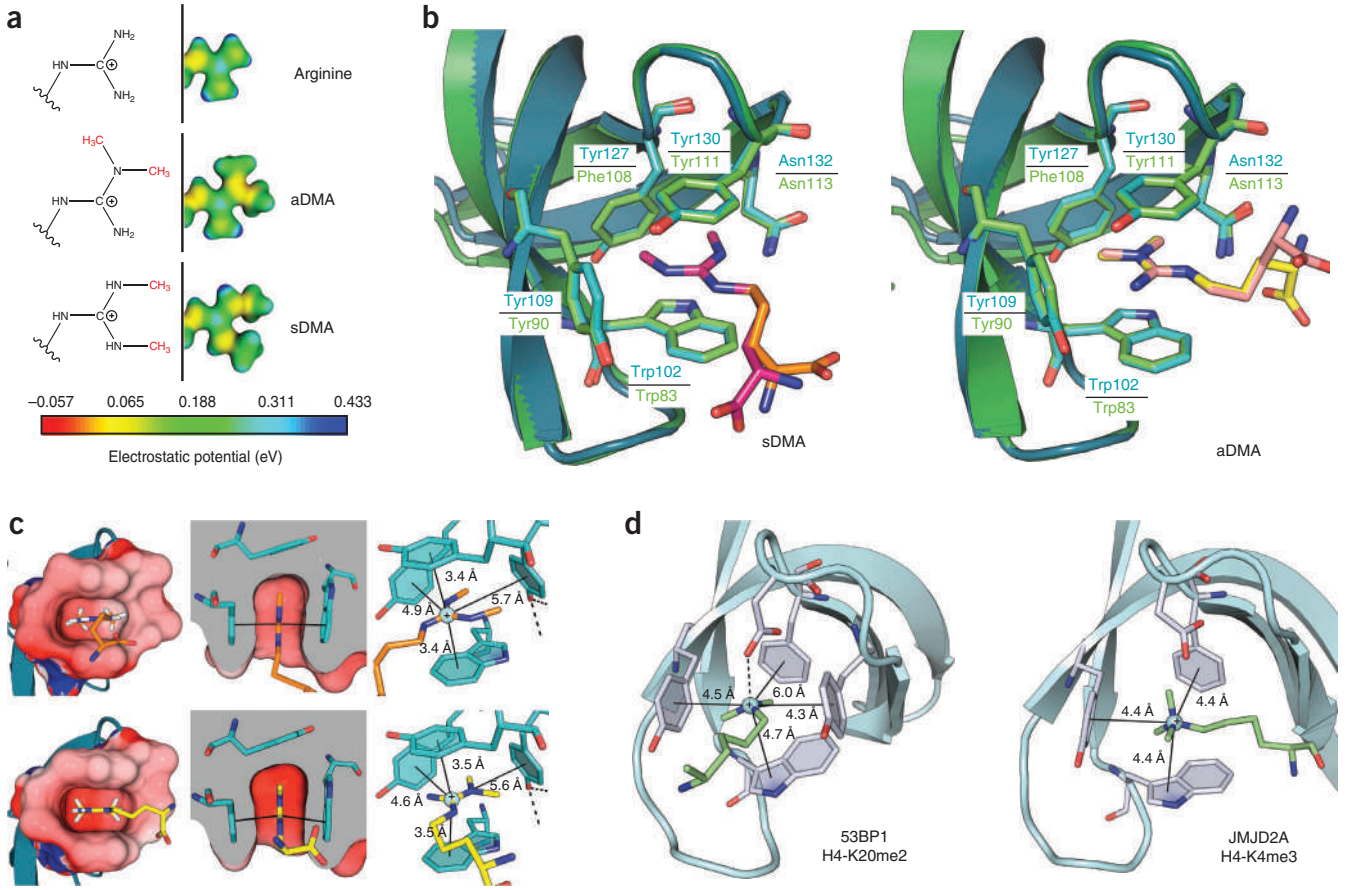


Figure 2 Recognition of sDMA and aDMA by SMN and SPF30 Tudor domains. (**a**) Molecular electrostatic potentials for arginine, aDMA and sDMA guanidino moieties (for details see Online Methods). (**b**) Overlay of SMN (blue) and SPF30 (green) structures in complex with sDMA (orange-purple with SMN-SPF30) and aDMA (yellow-pink with SMN-SPF30). The models represent the pairs with the lowest pairwise coordinate r.m.s. deviation (range 0.15–1.5 Å) from a (20 × 20) matrix of pairwise superpositions of the lowest-energy structures. (**c**) Aromatic cage electrostatics, stacked geometry and cation-π interactions for sDMA and aDMA recognition by SMN are shown at top and bottom, respectively. (**d**) Dimethyl- and trimethyllysine recognition by the Tudor domains of 53BP1 (PDB: 2IG0) and JMJD2A (PDB: 2GFA), respectively. For clarity only the Tudor domain that forms the aromatic cage and the methyllysine ligands are shown.

© 2011 Nature America, Inc. All rights reserved.

N132S reduced binding by 90% (**Supplementary Table 1b**). These data are consistent with the roles of interactions described above.

We examined binding in the TDRD2 protein (residues 327–420), which represents the trypophan-to-leucine substitution in our panel of aromatic cage mutations and has been structurally characterized[29]. Consistent with our structural analysis, no sDMA binding was observed for the TDRD2 Tudor domain by NMR or ITC (**Supplementary Fig. 8**), supporting the importance of this residue for DMA binding. Based on these observations, it can be predicted that about half of the Tudor domains in the TDRD family and the TUDOR *Drosophila* protein may not support efficient recognition of the dimethylarginine modification (**Fig. 3a** and **Supplementary Fig. 7**). Thus, it is likely that only a subset of Tudor domains in multi-TUDOR proteins mediates sDMA binding, whereas other Tudor domains may be involved in alternative interactions. Note that some members of the PHD finger and chromodomain families have also been shown to recognize nonmethylated ligands[26].

Design of Tudor-DMA binding specificity

A key difference between the aromatic cages of SMN and SPF30 is the aromatic residue at the backside of the cavity (SMN Tyr127 and SPF30 Phe108; **Figs. 1a** and **2b**). Tyrosine and phenylalanine are equally effective in cation-π binding, unless the hydroxyl group of the tyrosine is hydrogen bonded, in which case its cation-π binding strength increases substantially[27,28]. The structures of SMN Tudor bound to sDMA and aDMA revealed that the hydroxyl of the aromatic cage residue Tyr127 was engaged in a hydrogen bond triangle with the side chains of Glu134 and Gln136 (**Fig. 3b** and **Supplementary Fig. 9**), which is also present in the apo SMN Tudor domain[15,16]. To probe the role of this hydrogen bond for methylarginine binding, we substituted the corresponding residues and measured the binding affinities for the DMA ligands (**Supplementary Fig. 2**). Mutations that abolished the hydrogen bond triangle (SMN Y127F and E134S Q136V) attenuated binding to sDMA and aDMA to a similar extent (**Fig. 1d** and **Supplementary Table 1a**). On the other hand, introduction of a tyrosine at the corresponding position of SPF30 (F108Y) strengthened the interaction for both ligands. SPF30 already contains a glutamate (Glu113) at the position corresponding to Glu134 of SMN that can form a hydrogen bond to the tyrosine hydroxyl. Notably, the SPF30 F108Y T115Q double mutant strengthened the interaction to be comparable with that of SMN Tudor (**Fig. 1d** and **Supplementary Table 1a**), consistent with a fully completed hydrogen bond triangle (**Supplementary Fig. 9**).

SMA-linked point mutations in the SMN Tudor domain

Spinal muscular atrophy (SMA) is a neurodegenerative disease that results from mutations in the *SMN1* gene. Six pathogenic missense

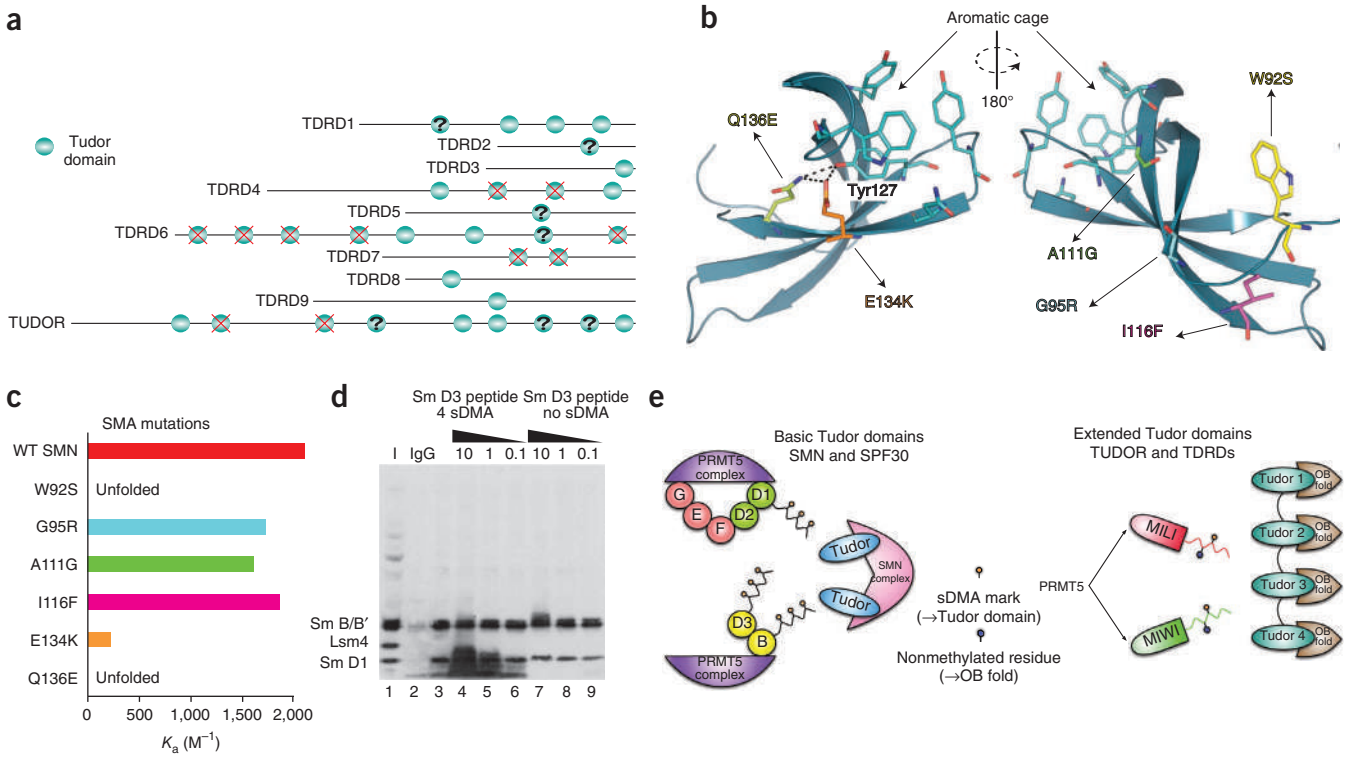


Figure 3 Implications of sDMA recognition in snRNP and piRNP assembly pathways. (**a**) Tudor domains in the sequences of mammalian TDRD proteins and TUDOR *Drosophila*. Question marks refer to Tudors with single substitutions, and red crosses refer to Tudors with two to four substitutions of aromatic cage residues. The latter are not expected to bind DMA. (**b**) Ribbon representation of SMN Tudor highlighting the side chains of SMA-related mutant proteins. The hydrogen bond triangle involving the side chains of Glu134 and Gln136 and the Tyr127 hydroxyl is indicated by a dashed line. (**c**) Association constants of wild-type and mutant SMN Tudors for sDMA binding (**Supplementary Table 1b**). (**d**) Total extracts from HeLa cells stably expressing SMN-GFP (green fluorescent protein) from an integrated bacterial artificial chromosome were analyzed by immunoprecipitation with anti-GFP, followed by western blotting with Y12 antibody that specifically recognizes sDMA marks. Lanes: (1) input, 5% of the total cell lysate; (2) proteins co-immunoprecipitated with nonimmune goat IgG; (3) proteins co-immunoprecipitated with SMN using anti-GFP and (4–9) incubated with the indicated concentrations (mM) of each peptide for competing the Tudor-sDMA interactions in the SMN complex. Note that the detection of additional low-molecular-weight bands in lanes 4–6 is due to the presence of the sDMA peptide, which reacts with Y12. (**e**) Possible roles for basic and extended Tudor domains in snRNP and piRNP pathways, respectively, as assembly platforms for ligands harboring sDMA modifications.

© 2011 Nature America, Inc. All rights reserved.

mutations in humans occur in the SMN Tudor domain[30–32]. We examined how these mutations affected the Tudor fold and the binding capacity of the protein to sDMA *in vitro* (**Fig. 3c** and **Supplementary Table 1b**). Two of the mutants (W92S and Q136E) were unfolded, as judged by the fingerprint NMR spectra (**Supplementary Fig. 10**) and did not bind sDMA. Notably, the E134K mutation reduced binding affinity for sDMA by an order of magnitude. Glu134, together with Gln136, participates in the hydrogen bond triangle with the aromatic cage Tyr127 hydroxyl (**Fig. 3b**). The glutamate to lysine substitution affected this hydrogen network and strongly reduced the sDMA interaction. Other SMA-linked single point mutations (G95R, A111G and I116F) perturbed the Tudor domain fold (**Supplementary Fig. 10**), but sDMA binding was similar to the wild-type protein (**Fig. 3c** and **Supplementary Table 1b**). These mutations may indirectly affect the protein stability and/or reduce the cellular levels of functional SMN protein in these patients. SMA mutants involving residues Glu134 (and presumably Trp92 and Gln136) strongly reduced sDMA binding by SMN Tudor, providing a possible link to the SMA phenotypes.

Role of SMN-sDMA recognition in snRNP assembly

Stepwise U snRNP assembly *in vitro* involves the association of Sm D3–B, D1–D2 and E–F–G hetero-oligomers onto the snRNA[33]. *In vivo*, U snRNP biogenesis depends on the PRMT5 and the SMN complexes[13,14,34]. First, the PRMT5 complex pre-organizes the Sm proteins and symmetrically dimethylates the arginine- and glycine-rich tails of Sm B/B′, D1 and D3. Subsequently, the SMN complex, which contains at least two copies of the SMN protein and several Gemin proteins, receives the Sm hetero-oligomers and promotes the formation of the heptameric Sm core of the U snRNPs. An array of Tudor domains thus presented by the SMN complex may serve as a platform for the recruitment of Sm subcomplexes by specific recognition of sDMA marks in the Sm tails (**Fig. 3d**,**e**). It has previously been shown that only sDMA-modified, but not unmodified, peptides representing the C-terminal arginine- and glycine-rich tails of Sm D1 and Sm D3 are able to pull down SMN from HeLa cell extracts[5]. Our co-immunoprecipitation experiments using overexpressed (**Fig. 3d**) or endogenous SMN protein (**Supplementary Fig. 11**) demonstrated that the Sm proteins present in the SMN complex were symmetrically dimethylated, supporting the role of the sDMA modification for the SMN interaction. Note that competition with the Sm D3 peptide, either unmodified or containing four sDMA residues, was not able to dissociate the methylated SMN partners from the SMN complex, indicating that the SMN complex is stabilized by various additional interactions. Indeed, a recently reported crystal structure has demonstrated that Gemin2 makes extensive contacts with an N-terminal helix of SMN and the pentamer of Sm proteins

(D1–D2–F–E–G)[35]. Although the sDMA modification of Sm proteins appears nonessential *in vitro*[34], the requirement for the PRMT5 complex[14,34,36] and the methylation status of the Sm proteins *in vivo* (**Fig. 3d** and **Supplementary Fig. 11**) argue that the readout of the sDMA modification by the SMN Tudor domain (**Fig. 2**) may regulate the kinetics and fidelity during early steps of ribonucleoprotein assembly processes *in vivo*.

DISCUSSION

Recently, crystal structures of extended Tudor domains (TUDOR *Drosophila* and SND1 proteins) in complex with symmetrically dimethylated peptides have been reported[22,23]. In these structures the basic Tudor fold is flanked by additional secondary structure elements. There are notable differences in the recognition of sDMA ligands by the extended Tudor domains compared to the prototypic Tudor domains of SMN and SPF30.

First, efficient binding by the extended Tudor domains requires additional contacts between the OB-fold extension and neighboring, non-methylated amino acids flanking the sDMA within the bound peptide[22,23]. This is in stark contrast to the SMN and SPF30 Tudor domains, where the basic Tudor fold alone is sufficient to recognize exclusively the sDMA marks in cognate peptides (**Supplementary Fig. 1**). Even though SMN and SPF30 show only the basic Tudor fold, they achieve low-micromolar affinities by external binding cooperativity, which results from the presence of multiple sDMA residues within a short stretch of ligand residues as presented by the Sm proteins (**Fig. 1** and **Supplementary Fig. 2**). This external binding cooperativity may compensate for the lack of additional interactions with residues adjacent to the sDMA mark seen in the complex structures of extended Tudor domains.

A second important difference between extended and basic Tudor domains concerns the conformation of the bound sDMA and the rotameric state of the asparagine residue that completes the binding pocket (**Supplementary Fig. 12**). In both reported extended Tudor domains[22,23], the asparagine side chain adopts the *gauche*(+) χ_1 rotameric state (**Supplementary Fig. 12**). In this arrangement the cavity accommodates the *anti-syn* conformation of sDMA, albeit with lower intrinsic affinity. In the basic Tudor domains of SMN and SPF30 presented here, the *gauche*(−) χ_1 rotameric state of the asparagine, which has been measured experimentally for the bound forms (**Supplementary Fig. 13**), enables binding to the *anti-anti* conformation of sDMA with higher intrinsic affinity (**Fig. 2**). This may additionally correlate with the tryptophan residue at the bottom of the cavity that is a distinct feature of SMN and SPF30 Tudor domains (**Supplementary Figs. 7** and **12**). Compared with the phenylalanine (or tyrosine in other instances) that occupies the equivalent position in the recently reported structures of extended Tudor domains[22,23], the tryptophan indole ring enhances the interaction with the methylarginine for two reasons: first, the six-membered ring of the tryptophan side chain mediates stronger π-cation interactions[27,28], and second, the connected six-and five-membered aromatic rings of the indole moiety provide a larger surface able to accommodate both methyl groups of the sDMA (if presented in an *anti-anti* conformation), thereby substantially increasing the van der Waals contacts[27,28].

Note that the *anti-syn* conformation of the sDMA is sterically incompatible with the asparagine *gauche*(−) χ_1 rotamer found in the basic Tudor domains and, vice versa, the asparagine *gauche*(+) χ_1 rotamer observed in the extended Tudor domains would exclude the *anti-anti* conformation of sDMA from the aromatic cage. Density function theory (DFT) calculations indicate that the ground-state energies for the two (free) sDMA conformations differ by 2.8 kcal mol^{-1}, with that of the *anti-syn* conformation being the most favorable. However, the activation energy for rotation about the C^{ζ}-N^{η} bond is 14 kcal mol^{-1} based on the DFT calculations (**Supplementary Fig. 13**). Thus, at room temperature[37], rotations about the C^{ζ}-N^{η} bond of the dimethylated guanidino group will allow for an (*anti-anti*) binding conformation that is distinct from the lowest-energy (*anti-syn*) conformation of the dimethylated arginine side chain when free in solution.

An important role for Tudor-sDMA interactions in RNP biogenesis has been demonstrated for the assembly of piRNA pathway components during germline development. The recognition of sDMA modifications in N-terminal regions of PIWI proteins (such as MILI and MIWI) by TDRD proteins is an interaction in the germline conserved by evolution[2,19–21]. However, TDRD proteins have acquired extensions in the Tudor fold accompanied by new intrinsic features in the recognition of methylarginines (**Supplementary Fig. 12**). The linkage of the Tudor domain to the OB fold has expanded the available interaction surface around the aromatic cage, offering the possibility of additional contacts with residues flanking the sDMA (**Fig. 3e**). Indeed, the OB fold is indispensable for the recognition of methylarginine peptides by the extended Tudor domains[22,23]. It not only assists the aromatic cage in the readout of the sDMA in the *anti-syn* conformation but also provides another level of selectivity for the peptide sequence flanking the modification. Given that TUDOR *Drosophila* and some TDRDs contain functional repeats of the extended Tudor domain, they may be capable of sorting out sDMA-containing effector proteins at specific points of the piRNA pathway (**Fig. 3e**).

Our structural findings show that the same general principles allow the basic Tudor domains of SMN and SPF30 to bind aDMA with subtle differences reflecting the slightly lower affinity for aDMA (**Supplementary Fig. 12**). These observations are consistent with a role for aDMA in Tudor-mediated interactions. For example, SMN interaction with the splicing factor CA150 is subject to regulation by aDMA modification[38]. In addition, the Tudor domain of TDRD3 selectively recognizes a peptide from the C-terminal domain of RNA polymerase II that is asymmetrically modified[39]. Finally, the *Drosophila* Vasa protein and its mouse homolog contain both sDMA and aDMA, but the sDMA modifications were found to be dispensable for the association with their Tudor binding partners[40]. Thus, it is possible that Vasa interaction with Tudor-containing partners relies on aDMA modification as well. The comparable binding affinities and similar modes of recognition of the SMN and SPF30 Tudor domain for sDMA and aDMA (**Fig. 2**) are consistent with a biological role of Tudor-aDMA interactions. Our structural analysis of the Tudor-aDMA interaction may serve as a model for understanding the interactions of Tudor-domain proteins with aDMA modifications in cognate ligands.

The tandem Tudor domains of 53BP1 (ref. 24) and the hybrid Tudor domains of JMJD2A (ref. 25) also recognize methyllysine residues through aromatic caging. However, the electrostatic interactions involving the positively charged methyllysine are weaker compared with dimethylarginine because of the longer distances to the rings of the aromatic cage (**Fig. 2d**). This is consistent with the observation that none of the histone methyllysine readers binds appreciably to the isolated methyllysine amino acid (K_d greater than millimolar)[26]. The same holds true for the extended Tudor domains of TUDOR *Drosophila* and SND1 proteins that bind sDMA in the *anti-syn* conformation[22,23]. The crystal structures indicate that the cation-π geometry is not optimal and the interaction between the cage and the sDMA headgroup may thus require further stabilization by a hydrogen bond (**Supplementary Fig. 12**). Considering the low affinities for individual methyllysine and methylarginine amino acids, different mechanisms have been devised to interpret these modifications. The Tudor domains that are fused to other domains—for example, to another Tudor (53BP1 and JMJD2A)

© 2011 Nature America, Inc. All rights reserved.

or to the OB fold (TUDOR *Drosophila* and SND1)—recognize methyllysines or methylarginines as part of a longer peptide sequence, where both the modification and the flanking residues contribute to the binding affinity and specificity[22–26]. In contrast, the basic Tudor domains of SMN and SPF30 achieve increased affinities for arginine- and glycine-rich peptides by external binding cooperativity linked to the presence of multiple DMA modifications (**Fig. 1c** and **Supplementary Table 1a**) and may represent the prototypic mode of Tudor-DMA recognition that acquired new intrinsic properties in the extended architecture of the Tudor fold.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/nsmb/.

Accession codes. The atomic coordinates for the NMR ensembles have been deposited in the Protein Data Bank under accession numbers 4A4E (SMN–sDMA), 4A4F (SPF30–sDMA), 4A4G (SMN–aDMA) and 4A4H (SPF30–aDMA). The chemical shift assignments have been deposited in the Biological Magnetic Resonance Data Bank under accession numbers 18005 (SMN–sDMA), 18006 (SPF30–sDMA), 18007 (SMN–aDMA) and 18008 (SPF30–aDMA).

Note: Supplementary information is available on the Nature Structural & Molecular Biology website.

ACKNOWLEDGMENTS

The authors are grateful to G. Demiraslan (Helmholtz Zentrum München) for protein production. We thank I. Poser and T. Hyman (Max Planck Institute of Molecular Cell Biology and Genetics) for the gift of the SMN-GFP stable cell line, J. Brennecke for helpful discussion and the Bavarian NMR Centre for NMR time. K.T. acknowledges support by the Alexander von Humboldt foundation; T.M. acknowledges support by a European Molecular Biology Organization Long Term Fellowship and a Schrödinger Fellowship from the Austrian Science Fund. This work was supported by the Deutsche Forschungsgemeinschaft (M.S. and K.M.N.).

AUTHOR CONTRIBUTIONS

K.T. carried out molecular biology experiments, protein purification, NMR analysis and structure calculations. T.M. did quantum chemical calculations and contributed to NMR experiments. D.F. analyzed the ITC experiments. K.M.N. and U.F. designed, and M.M. and C.E. conducted, immunoprecipitation experiments. K.T. and M.S. conceived and designed the project and wrote the paper. All authors discussed the results and commented on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/nsmb/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Bedford, M.T. & Clarke, S.G. Protein arginine methylation in mammals: who, what, and why. *Mol. Cell* **33**, 1–13 (2009).
2. Siomi, M.C., Mannen, T. & Siomi, H. How does the royal family of Tudor rule the PIWI-interacting RNA pathway? *Genes Dev.* **24**, 636–646 (2010).
3. Boisvert, F.M., Cote, J., Boulanger, M.C. & Richard, S. A proteomic analysis of arginine-methylated protein complexes. *Mol. Cell. Proteomics* **2**, 1319–1330 (2003).
4. Brahms, H., Meheus, L., de Brabandere, V., Fischer, U. & Luhrmann, R. Symmetrical dimethylation of arginine residues in spliceosomal Sm protein B/B′ and the Sm-like protein LSm4, and their interaction with the SMN protein. *RNA* **7**, 1531–1542 (2001).
5. Friesen, W.J., Massenet, S., Paushkin, S., Wyce, A. & Dreyfuss, G. SMN, the product of the spinal muscular atrophy gene, binds preferentially to dimethylarginine-containing protein targets. *Mol. Cell* **7**, 1111–1117 (2001).
6. Côté, J. & Richard, S. Tudor domains bind symmetrical dimethylated arginines. *J. Biol. Chem.* **280**, 28476–28483 (2005).
7. Friberg, A., Corsini, L., Mourao, A. & Sattler, M. Structure and ligand binding of the extended Tudor domain of *D. melanogaster* Tudor-SN. *J. Mol. Biol.* **387**, 921–934 (2009).
8. Bühler, D., Raker, V., Lührmann, R. & Fischer, U. Essential role for the Tudor domain of SMN in spliceosomal U snRNP assembly: implications for spinal muscular atrophy. *Hum. Mol. Genet.* **8**, 2351–2357 (1999).
9. Pellizzoni, L., Kataoka, N., Charroux, B. & Dreyfuss, G. A novel function for SMN, the spinal muscular atrophy disease gene product, in pre-mRNA splicing. *Cell* **95**, 615–624 (1998).
10. Pellizzoni, L., Yong, J. & Dreyfuss, G. Essential role for the SMN complex in the specificity of snRNP assembly. *Science* **298**, 1775–1779 (2002).
11. Meister, G. *et al.* SMNrp is an essential pre-mRNA splicing factor required for the formation of the mature spliceosome. *EMBO J.* **20**, 2304–2314 (2001).
12. Rappsilber, J., Ajuh, P., Lamond, A.I. & Mann, M. SPF30 is an essential human splicing factor required for assembly of the U4/U5/U6 tri-small nuclear ribonucleoprotein into the spliceosome. *J. Biol. Chem.* **276**, 31142–31150 (2001).
13. Paushkin, S., Gubitz, A.K., Massenet, S. & Dreyfuss, G. The SMN complex, an assemblyosome of ribonucleoproteins. *Curr. Opin. Cell Biol.* **14**, 305–312 (2002).
14. Meister, G. & Fischer, U. Assisted RNP assembly: SMN and PRMT5 complexes cooperate in the formation of spliceosomal UsnRNPs. *EMBO J.* **21**, 5853–5863 (2002).
15. Selenko, P. *et al.* SMN Tudor domain structure and its interaction with the Sm proteins. *Nat. Struct. Biol.* **8**, 27–31 (2001).
16. Sprangers, R., Groves, M.R., Sinning, I. & Sattler, M. High-resolution X-ray and NMR structures of the SMN Tudor domain: conformational variation in the binding site for symmetrically dimethylated arginine residues. *J. Mol. Biol.* **327**, 507–520 (2003).
17. Hebert, M.D., Shpargel, K.B., Ospina, J.K., Tucker, K.E. & Matera, A.G. Coilin methylation regulates nuclear body formation. *Dev. Cell* **3**, 329–337 (2002).
18. Renvoisé, B. *et al.* Distinct domains of the spinal muscular atrophy protein SMN are required for targeting to Cajal bodies in mammalian cells. *J. Cell Sci.* **119**, 680–692 (2006).
19. Vagin, V.V. *et al.* Proteomic analysis of murine Piwi proteins reveals a role for arginine methylation in specifying interaction with Tudor family members. *Genes Dev.* **23**, 1749–1762 (2009).
20. Kirino, Y. *et al.* Arginine methylation of Piwi proteins catalysed by dPRMT5 is required for Ago3 and Aub stability. *Nat. Cell Biol.* **11**, 652–658 (2009).
21. Nishida, K.M. *et al.* Functional involvement of Tudor and dPRMT5 in the piRNA processing pathway in *Drosophila* germlines. *EMBO J.* **28**, 3820–3831 (2009).
22. Liu, H. *et al.* Structural basis for methylarginine-dependent recognition of Aubergine by Tudor. *Genes Dev.* **24**, 1876–1881 (2010).
23. Liu, K. *et al.* Structural basis for recognition of arginine methylated Piwi proteins by the extended Tudor domain. *Proc. Natl. Acad. Sci. USA* **107**, 18398–18403 (2010).
24. Botuyan, M.V. *et al.* Structural basis for the methylation state–specific recognition of histone H4–K20 by 53BP1 and Crb2 in DNA repair. *Cell* **127**, 1361–1373 (2006).
25. Huang, Y., Fang, J., Bedford, M.T., Zhang, Y. & Xu, R.M. Recognition of histone H3 lysine-4 methylation by the double Tudor domain of JMJD2A. *Science* **312**, 748–751 (2006).
26. Taverna, S.D., Li, H., Ruthenburg, A.J., Allis, C.D. & Patel, D.J. How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nat. Struct. Mol. Biol.* **14**, 1025–1040 (2007).
27. Mecozzi, S., West, A.P. Jr. & Dougherty, D.A. Cation-π interactions in aromatics of biological and medicinal interest: electrostatic potential surfaces as a useful qualitative guide. *Proc. Natl. Acad. Sci. USA* **93**, 10566–10571 (1996).
28. Gallivan, J.P. & Dougherty, D.A. Cation-π interactions in structural biology. *Proc. Natl. Acad. Sci. USA* **96**, 9459–9464 (1999).
29. Chen, C. *et al.* Mouse Piwi interactome identifies binding mechanism of Tdrkh Tudor domain to arginine methylated Miwi. *Proc. Natl. Acad. Sci. USA* **106**, 20336–20341 (2009).
30. Chari, A., Paknia, E. & Fischer, U. The role of RNP biogenesis in spinal muscular atrophy. *Curr. Opin. Cell Biol.* **21**, 387–393 (2009).
31. Pellizzoni, L., Charroux, B. & Dreyfuss, G. SMN mutants of spinal muscular atrophy patients are defective in binding to snRNP proteins. *Proc. Natl. Acad. Sci. USA* **96**, 11167–11172 (1999).
32. Eggert, C., Chari, A., Laggerbauer, B. & Fischer, U. Spinal muscular atrophy: the RNP connection. *Trends Mol. Med.* **12**, 113–121 (2006).
33. Raker, V.A., Plessel, G. & Lührmann, R. The snRNP core assembly pathway: identification of stable core protein heteromeric complexes and an snRNP subcore particle *in vitro*. *EMBO J.* **15**, 2256–2269 (1996).
34. Chari, A. *et al.* An assembly chaperone collaborates with the SMN complex to generate spliceosomal SnRNPs. *Cell* **135**, 497–509 (2008).
35. Zhang, R. *et al.* Structure of a key intermediate of the SMN complex reveals gemin2's crucial function in snRNP assembly. *Cell* **146**, 384–395 (2011).
36. Friesen, W.J. *et al.* The methylosome, a 20S complex containing JBP1 and pICln, produces dimethylarginine-modified Sm proteins. *Mol. Cell. Biol.* **21**, 8289–8300 (2001).
37. Kessler, H. Detection of hindered rotation and inversion by NMR spectroscopy. *Angew. Chem. Int. Ed. Engl.* **9**, 219–235 (1970).
38. Cheng, D., Cote, J., Shaaban, S. & Bedford, M.T. The arginine methyltransferase CARM1 regulates the coupling of transcription and mRNA processing. *Mol. Cell* **25**, 71–83 (2007).
39. Sims, R.J. *et al.* The C-terminal domain of RNA polymerase IIis modified by site-specific methylation. *Science* **332**, 99–103 (2011).
40. Kirino, Y. *et al.* Arginine methylation of Vasa protein is conserved across phyla. *J. Biol. Chem.* **285**, 8148–8154 (2010).

© 2011 Nature America, Inc. All rights reserved.

ONLINE METHODS

Sample preparation. The plasmids expressing $SMN_{84–147}$ and $SPF30_{65–128}$ with N-terminal fused histidine tags were transformed into *Escherichia coli* BL21 (DE3) cells. All mutants were constructed using site-directed mutagenesis, and the clones were confirmed by sequencing. Cells were grown at 20 °C in the presence of kanamycin. ^{13}C- and ^{15}N-labeled samples were prepared by growing cells in minimal medium supplemented with $^{15}NH_4Cl$ (1g liter^{-1}) and $^{13}C_6$ glucose (2g liter^{-1}) as the sole nitrogen and carbon sources, respectively. Protein synthesis was induced by the addition of 0.5 mM of isopropyl-1-thio-D-galactopyranoside (IPTG) at OD_{600} ~0.8, and cells were collected after 20–24 h. Cells were resuspended in lysis buffer containing 50 mM Tris-HCl (pH 8.0), 500 mM NaCl, 10 mM imidazole, 1 mM tris-(2-carboxyethyl)-phosphine (TCEP), 0.1% (v/v) Triton X-100 , lysed by sonication and centrifuged at 15,000*g*. Two purification steps were used for all protein samples. The first one involved Ni-NTA gravity-flow resin (Qiagen) followed by removal of the tag, and the second one used a Superdex-75 size exclusion column (GE Healthcare). MMA, aDMA and sDMA were purchased from Enzo Life Sciences and sDMA-containing peptides from Peptide Specialty Laboratories.

NMR spectroscopy. The samples used for structure determination contained 4 mM protein (SMN or SPF30) and 18 mM of the respective ligand (aDMA or sDMA) in 20 mM sodium phosphate (pH 6.5), 50 mM NaCl with 7% (v/v) ^2H_2O added for the lock. Spectra were recorded at 298 K using AVIII600, AVIII750 or AV900 Bruker NMR spectrometers equipped with cryogenic triple resonance gradient probes (except AVIII750). Spectra were processed using NMRPipe[41]. Protein backbone and side chain assignments were obtained from standard triple resonance experiments[42]. For every complex structure six NOESY experiments were acquired. NOE distance restraints were obtained from ^{15}N- and ^{13}C-edited 3D NOESY spectra (mixing time 70 ms). Intermolecular NOE restraints were derived from 3D ^{15}N,^{13}C-filtered, ^{15}N- and ^{13}C-edited NOESY spectra[43] (mixing time 70 ms). Stereospecific assignments of the methyl groups of valine and leucine were obtained using 10% ^{13}C-labeled protein samples[44]. Quantitative J correlation spectroscopy measurements to determine the χ_1 rotamer of the asparagine residue have been carried out as described[45,46].

To confirm the bound conformation of the ligands based on NOEs derived from samples with the ligands in excess, we measured intermolecular NOEs using samples with the protein in excess (**Fig. 1f**). In these samples, where the chemical shifts of the ligands represent the genuine bound form, ^{15}N,^{13}C-filtered 2D and 3D ^{15}N- and ^{13}C-edited NOESY spectra were recorded at different temperatures (from 278 to 318 K) and different mixing times (from 10 to 200 ms) to shift the exchange regime in order to obtain a comprehensive qualitative NOE analysis.

Structure determination. Automated NOESY cross-peak assignments and structure calculations with torsion angle dynamics were done by using the program CYANA 3.0 (ref. 47). For every single peak list, the distances between the protons were calibrated using CYANA's iterative calibration algorithm. The algorithm employs a $V = A\,u^{-6}$ relationship between the peak volume, V and the corresponding upper distance bound, u. The constant A is assumed such that the median of all peak volumes in a given NOESY spectrum corresponds to a distance of 4 Å and accordingly calibrates the distances for all the cross peaks (for a range between 2.4 and 6 Å). The set of NOE distance restraints derived from CYANA together with φ and ψ backbone dihedral angle restraints derived from TALOS$^+$ (ref. 48) based on the chemical shifts were used for water refinement[49] using CNS[50]. Ligands have been parametrized by using the Dundee PRODRG2 Server[51], applying the crystal structures from the Heterocompound Information Centre (Uppsala, Sweden)[52]. The compound names are DA2 for aDMA and 2MR for sDMA. All structures were validated by using iCing (http://nmr.cmbi.ru.nl/icing/). Ramachandran plot statistics for residues in most favored regions, additional allowed regions, generously allowed regions, disallowed regions are: SMN–sDMA 89.5%, 10.1%, 0.2%, 0.3%; SPF30–sDMA 96.7%, 2.9%, 0.3%, 0.2%; SMN–aDMA 87.9%, 11.8%, 0.2%, 0.1%; SPF30–aDMA 95.6%, 4.4%, 0%, 0%. Molecular images were generated with PyMol (Schrödinger) and Molekel (http://molekel.cscs.ch/wiki/pmwiki.php).

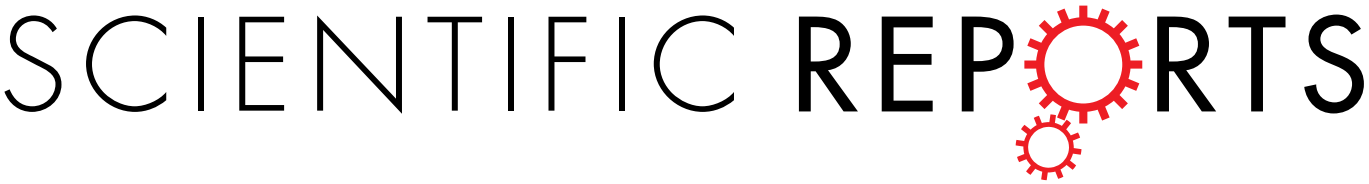
***Ab initio* ligand calculations.** Arginine, aDMA and sDMA starting geometries were used in the Gaussian 03 program suite (Gaussian), using Becke's three-parameter hybrid Hartree-Fock density functional method[53] in combination with the Lee-Yang-Parr correlation functional (B3LYP)[54] and the 6–31g(d) basis set. All structures were optimized and characterized by frequency calculations as true minima. Atomic charges were computed from the Mulliken population analysis. Molecular electrostatic potentials for the arginine, aDMA and sDMA guanidino moieties in **Figure 2a** are plotted in units of eV on the electron isodensity surface at $\rho = 0.05\ e$ Å^{-3}; "e" is the electron charge.

ITC experiments. Calorimetric titrations were carried out on an iTC200 microcalorimeter (MicroCal) at 25 °C. Protein and ligand samples were dissolved in 20 mM sodium phosphate (pH 6.5), 50 mM NaCl. The 200-µl sample cell was filled with a 0.5 or 1 mM solution of protein, and the 40-µl injection syringe with 20 or 40 mM of the titrating ligand. Each titration consisted of a preliminary 0.2-µl injection followed by 40 subsequent 1-µl injections. The heat of the injections was corrected for the heat of dilution of every ligand into the buffer. At least two replicas were done for each experiment. Binding thermodynamic models[55,56] were tested to interpret the calorimetric data. The fit attempts were accomplished using the nonlinear Levenberg-Marquardt method[57]. The errors of each fitting parameter were calculated with a 95.4% confidence limit by the Monte Carlo simulation method. An eventual error of 10% in the protein and/or ligand concentration was also taken into account. In all cases errors were within 10% of the reported values.

Immunoprecipitation. SMN complexes were immunoprecipitated with anti-GFP antibody from HeLa cells stably expressing SMN-GFP from a bacterial artificial chromosome and washed with NET-2 buffer (50 mM Tris pH 7.5, 150 mM NaCl, 0.05% (v/v) NP-40). Complexes were then incubated twice for 15 min in NET-2 buffer containing Sm D3 peptide (Peptide Specialty Laboratories), with or without 4 sDMA modifications, and washed after each incubation. Remaining complexes were eluted by boiling in Laemmli buffer, followed by western blotting with Y12 antibody.

41. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).
42. Sattler, M., Schleucher, J.R. & Griesinger, C. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nucl. Magn. Reson. Spectrosc.* **34**, 93–158 (1999).
43. Breeze, A.L. Isotope-filtered NMR methods for the study of biomolecular structure and interactions. *Prog. Nucl. Magn. Reson. Spectrosc.* **36**, 323–372 (2000).
44. Neri, D., Szyperski, T., Otting, G., Senn, H. & Wuthrich, K. Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 repressor by biosynthetically directed fractional ^{13}C labeling. *Biochemistry* **28**, 7510–7516 (1989).
45. Grzesiek, S. & Bax, A. A three-dimensional NMR experiment with improved sensitivity for carbonyl-carbonyl J correlation in proteins. *J. Biomol. NMR* **9**, 207–211 (1997).
46. Hu, J.S. & Bax, A. χ_1 angle information from a simple two-dimensional NMR experiment that identifies trans $^3J_{NC\gamma}$ couplings in isotopically enriched proteins. *J. Biomol. NMR* **9**, 323–328 (1997).
47. Güntert, P. Automated structure determination from NMR spectra. *Eur. Biophys. J.* **38**, 129–143 (2009).
48. Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. TALOS$^+$: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR* **44**, 213–223 (2009).
49. Linge, J.P., Williams, M.A., Spronk, C.A., Bonvin, A.M. & Nilges, M. Refinement of protein structures in explicit solvent. *Proteins* **50**, 496–506 (2003).
50. Brünger, A.T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54**, 905–921 (1998).
51. Schüttelkopf, A.W. & van Aalten, D.M. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 1355–1363 (2004).
52. Kleywegt, G.J. Crystallographic refinement of ligand complexes. *Acta Crystallogr. D Biol. Crystallogr.* **63**, 94–100 (2007).
53. Becke, A. Density-functional thermochemistry. III. the role of exact exchange. *J. Chem. Phys.* **98**, 5648–5652 (1993).
54. Lee, C., Yang, W. & Parr, R.G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B Condens. Matter.* **37**, 785–789 (1988).
55. Gill, S.J. Thermodynamics of ligand-binding to proteins. *Pure Appl. Chem.* **61**, 1009–1020 (1989).
56. Capaldi, S. *et al.* The X-ray structure of zebrafish (*Danio rerio*) ileal bile acid-binding protein reveals the presence of binding sites on the surface of the protein molecule. *J. Mol. Biol.* **385**, 99–116 (2009).
57. Press, W., Teukolsky, S., Vetterling, W. & Flannery, B. *Numerical Recipes 3rd Edition: The Art of Scientific Computing* (Cambridge University Press, 2007).

© 2011 Nature America, Inc. All rights reserved.



OPEN

The redox environment triggers conformational changes and aggregation of hIAPP in Type II Diabetes

Received: 23 September 2016
Accepted: 03 February 2017
Published: 13 March 2017

Diana C. Rodriguez Camargo[1,2], Konstantinos Tripsianes[3], Katalin Buday[1], Andras Franko[1,4], Christoph Göbl[1,2], Christoph Hartlmüller[2], Riddhiman Sarkar[1,2], Michaela Aichler[1], Gabriele Mettenleiter[1], Michael Schulz[1], Annett Böddrich[5], Christian Erck[6], Henrik Martens[6], Axel Karl Walch[1], Tobias Madl[1,2,7], Erich E. Wanker[5], Marcus Conrad[1], Martin Hrabě de Angelis[1,4,8] & Bernd Reif[1,2]

Type II diabetes (T2D) is characterized by diminished insulin production and resistance of cells to insulin. Among others, endoplasmic reticulum (ER) stress is a principal factor contributing to T2D and induces a shift towards a more reducing cellular environment. At the same time, peripheral insulin resistance triggers the over-production of regulatory hormones such as insulin and human islet amyloid polypeptide (hIAPP). We show that the differential aggregation of reduced and oxidized hIAPP assists to maintain the redox equilibrium by restoring redox equivalents. Aggregation thus induces redox balancing which can assist initially to counteract ER stress. Failure of the protein degradation machinery might finally result in β-cell disruption and cell death. We further present a structural characterization of hIAPP in solution, demonstrating that the N-terminus of the oxidized peptide has a high propensity to form an α-helical structure which is lacking in the reduced state of hIAPP. In healthy cells, this residual structure prevents the conversion into amyloidogenic aggregates.

Type II diabetes (T2D) is the most common form of diabetes. The disease is caused by a combination of genetic and environmental factors, and results in insulin resistance and β-cell failure[1]. T2D is also linked to the tissue inflammation response[2]. The exact patho-mechanism of the disease is, however, not understood so far. In addition to insulin resistance, lipotoxicity, endoplasmic reticulum (ER) and oxidative stress, as well as amyloid deposition have been described as secondary effects of the disease[2]. The amyloid plaques formed within the β-cells of the pancreatic islets of Langerhans[3] are composed of the 37 amino acid polypeptide human Islet Amyloid Polypeptide (hIAPP). In recent studies it has been shown that soluble oligomers of this hormone are responsible for cell toxicity[4,5]. Yet until now, it is not understood how hIAPP changes from a functional into a pathophysiological form. *In vivo*, hIAPP is soluble at concentrations of 1–4 mM[6–8], while *in vitro* hIAPP aggregates at 1000 times lower concentrations. Protein aggregation is influenced by concentration and pH[9]. In T2D, the redox state of the ER is shifted towards more reducing conditions, resulting in protein secretion and folding deficiencies[10–12]. So far, it is unclear how this change of the redox environment affects hIAPP aggregation. We show here that the redox conditions play an important role in the aggregation of hIAPP. Furthermore, we present the first detailed structural analysis of monomeric hIAPP in aqueous buffer. We find that the disulfide bridge in $hIAPP^{ox}$ stabilizes an α-helical structure at the N-terminus of the peptide, and thus protects the peptide from aggregation.

[1]Helmholtz Zentrum München, Ingolstädter Landstr. 1, Neuherberg 85764, Germany. [2]Munich Center for Integrated Protein Science (CIPS-M) at Department Chemie, Technische Universität München (TUM), Germany. [3]Central European Institute of Technology (CEITEC), Masaryk University, Kamenice 5, Brno 62500, Czech Republic. [4]German Center for Diabetes Research (DZD e.V.), Neuherberg 85764, Germany. [5]Max-Delbrück-Center Berlin (MDC), Robert-Rössle-Str. 10, Berlin 13125, Germany. [6]Synaptic Systems GmbH, Rudolf-Wissell-Straße 28, Göttingen, 37079, Germany. [7]Institute of Molecular Biology & Biochemistry, Center of Molecular Medicine, Medical University of Graz, Austria. [8]Technische Universität München, Center of Life and Food Sciences Weihenstephan, Freising 85354, Germany. Correspondence and requests for materials should be addressed to B.R. (email: reif@tum.de)

Results

Assembly Kinetics of hIAPP. To better understand the influence of the redox environment on the aggregation kinetics of hIAPP, we first performed Thioflavin T (ThT) assays. In a cell, the ratio of reduced glutathione (GSH) to oxidized glutathione (GSSG) is often employed as a measure of cellular oxidative stress. In order to induce a particular oxidative state of hIAPP, we added the redox couple GSH/GSSG to the peptide solution (Fig. 1A,B). A sigmoidal fit of the ThT results yields the time at which the half-maximum fluorescence intensity has been obtained (T_{50}). We found that the aggregation kinetics of hIAPP are affected by the oxidative environment. $hIAPP^{red}$ aggregates faster in comparison to $hIAPP^{ox}$. However, mixtures of $hIAPP^{red}$ and $hIAPP^{ox}$ yield the same rate of aggregation as $hIAPP^{red}$. Within the experimental error, we found the same T_{50} values for redox pairs that yield (100:0)%, (80:20)% or (50:50)% reduced and oxidized hIAPP. This suggests that the aggregation of hIAPP is controlled by the reduced form. The ThT results are supported by NMR experiments which show that $hIAPP^{red}$ has a 3-times faster aggregation kinetics in comparison to $hIAPP^{ox}$ (Fig. 1C). To validate the biological significance of the differential aggregation behaviour of $hIAPP^{ox}$ and $hIAPP^{red}$, we determined the redox potential for the formation of the disulfide bond in hIAPP using NMR spectroscopy. $^{1}H,^{15}N$ HSQC spectra were recorded after incubating the protein with varying ratios of reduced and oxidized glutathione. A plot of the average peak intensities as a function of the $[GSH]^2/[GSSG]$ ratio is shown in Fig. 1D. As a secondary abscissa, the calculated redox potential is represented. Using the Nernst equation with the respective correction for temperature and pH[13], we calculated a redox potential of (-129 ± 4) mV for hIAPP. The ER has an electrochemical potential that varies between -231 to -160 mV, and is dependent on the employed method[14]. Under these conditions, less than 30% of the peptide exists in the oxidized state. This theoretical redox state assumes quasi-equilibration with the redox couple GSH/GSSG.

To analyze the hIAPP oxidation state *in vivo*, we used a transgenic mouse model (FVB/N-Tg(Ins2-IAPP) RHFSoel/J) in which the human IAPP sequence has been incorporated into the mouse genome on chromosome 15 (https://www.jax.org/strain/008232). We found that the morphology of the islets is dramatically changed for transgenic mice that harbor two human gene copies (TG/TG) (Fig. 1F, left) compared to healthy controls. We employed a fibril specific antibody to study whether amyloid fibrils participate in this process. We found amyloid fibrils in insulin producing β-cells in TG/TG animals (Fig. 1F). These were not present in healthy animals. These results suggest that amyloid fibrils in TG/TG mice potentially induce impaired islet morphology. To quantify the amount of reduced or oxidized hIAPP *in vivo*, we added N-ethylmaleimide (NEM) to TG/TG and +/+ mouse islets to yield an alkylation of free sulfhydryl groups. NEM has the ability to cross cellular membranes and thus does not require cell lysis. Modified $hIAPP^{red}$-NEM yielded an increased molecular weight which could be detected by SDS-PAGE. Soluble oligomeric hIAPP could be identified using the hIAPP specific antibody A133[15]. In TG/TG islets, both reduced and oxidized hIAPP were found (Fig. 1E, right). As expected, no aggregates were found in the islets of the+/+control mice (Fig. 1E, right). As a reference, an aggregated sample of recombinant, oxidized hIAPP was investigated as well (Fig. 1E, left). We observed that samples that were taken at a later time point yielded a more disperse distribution of oligomeric states. An incubation time of 72 h for an oxidized hIAPP sample produced a gel band at approximately 60 kDa, which corresponds in molecular weight to the band of the hIAPP sample extracted from the diabetes mouse model islets (Fig. 1E, right, last sample).

In order to probe if the differential aggregation behavior results in a continuous conversion of $hIAPP^{ox}$ into $hIAPP^{red}$ at defined redox conditions, electron microscopic (EM) images of hIAPP aggregates that are obtained in buffers with different amounts of GSH and GSSG were recorded (Fig. 2A–C). Under conditions, in which the redox equilibrium is shifted to produce a defined amount of $hIAPP^{ox}$ (50:50 = $hIAPP^{red}$:$hIAPP^{ox}$) (Fig. 2C), the fibril morphology resembles the morphology obtained for the pure $hIAPP^{red}$ (Fig. 2B). By contrast, pure $hIAPP^{ox}$ (Fig. 2A) showed a different fibril morphology. To confirm this observation, we carried out solution-state HSQC NMR experiments of hIAPP aggregates that were dissolved in DMSO prior to the experiment. These experiments allow to probe the hIAPP redox state of the hIAPP aggregates. Again, we found that the hIAPP spectra of the samples in which the redox properties in the buffer has been adjusted to yield 100% $hIAPP^{red}$ or a 1:1 ratio of $hIAPP^{ox}$ and $hIAPP^{red}$ are rather similar, whereas the sample containing pure $hIAPP^{ox}$ looked significantly different (Fig. 2E). This indicates that aggregation in fact shifts the equilibrium to the reduced state of the peptide. Based on these results, we suggest a model for hIAPP aggregation in the cell that is dependent on its particular redox environment (Fig. 2F). External factors such as oxidative stress or ER stress, hyperglycemia, lipotoxity, or genetic factors have an impact on the ER redox state. A small shift of the redox conditions can result in generation of a small amount of $hIAPP^{red}$. $hIAPP^{red}$ preferentially aggregates and is depleted from the pool of hIAPP molecules in solution. As the redox state in the cell is buffered, more $hIAPP^{ox}$ is transformed to its reduced form which again has a higher propensity to form fibrils. Eventually, all available hIAPP is converted into its reduced state which is finally deposited in insoluble aggregates.

Structural analysis of hIAPP. To study the mechanism of hIAPP aggregation, the conformation of the peptide was determined under physiological conditions. Recombinant C-terminal amidated hIAPP(1-37) was produced using the protocol published recently[17]. hIAPP has a strong tendency to aggregate in aqueous buffer. However, low pH (pH 5.3), temperature (4 °C) and concentration (100 μM) reduced the rate of fibril formation. Under these conditions, hIAPP(1-37) yielded a CD spectrum which contained contributions from α-helix and random coil secondary structure (Fig. 3A). Comparison of the oxidized and the reduced form of the peptide suggests that a loss of the disulfide bond induces a loss of the α-helical structure. To support this observation, we recorded solution-state NMR experiments. $^{1}H,^{15}N$ HSQC spectra of reduced and oxidized hIAPP yielded significant chemical shift differences, in particular for Cys-2 and Cys-7 as well as for residues in the vicinity of the disulfide bond (Fig. 2D). Chemical shift assignments have been obtained from 3D HNCACB experiments[18]. The analysis of 13 $C^{\alpha/\beta}$ secondary chemical shifts reveals a strong helical propensity for the N-terminal half of the oxidized hIAPP (residues 5-17), and a random coil conformation for the C-terminal half of the peptide (Fig. 3E).

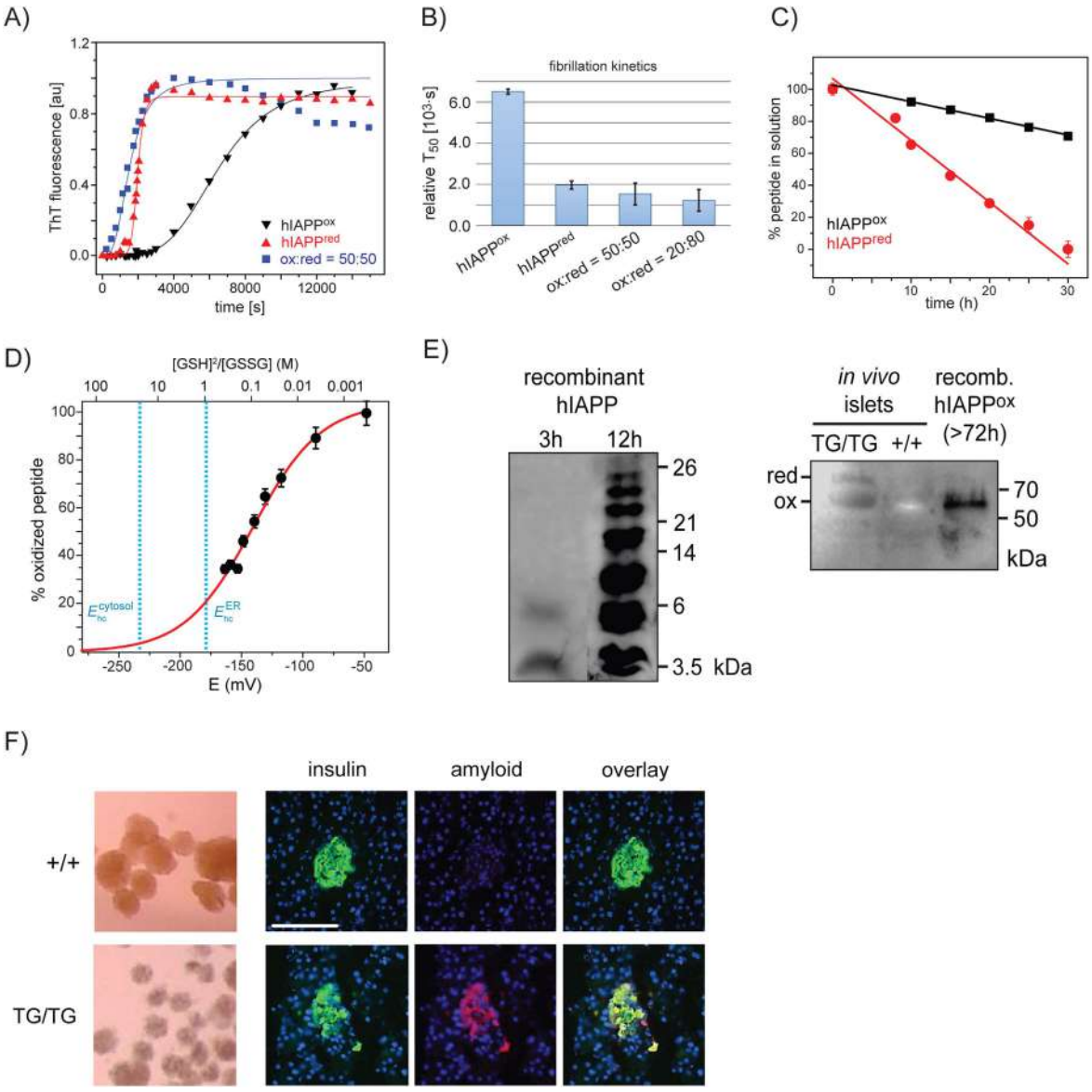


Figure 1. (**A**) Thioflavin T (ThT) assay using a solution of 35 μM of hIAPP, pH 5.3 and 30 °C. Experiments were performed by adding defined amounts of GSH and GSSG to yield varying percentages of reduced and oxidized hIAPP in solution. In all three curves, ThT fluorescence intensities are normalized to 1. ThT fluoresence curves with absolute intensities are represented in Fig. S2. (**B**) T_{50} values for hIAPP aggregation for different redox conditions. T_{50} corresponds to the time that is required to reach half-maximum ThT fluoresence intensity. (**C**) Aggregation kinetics of $hIAPP^{ox}$ and $hIAPP^{red}$ measured by solution-state NMR. $hIAPP^{red}$ aggregates significantly faster in comparison to $hIAPP^{ox}$. The intensities of the NMR resonances directly indicate the amount of soluble $hIAPP^{ox}$ and $hIAPP^{red}$. Both peptides were solubilized at a concentation of 100 μM in aqueous buffer (30 mM acetic acid, pH 5.3, 4 °C) (**D**) Determination of the electrochemical potential of hIAPP. The relative amount of $hIAPP^{ox}$ is represented as a function of the concentrations of GSH and GSSG. For the analysis, the peak intensity of Cys-7 has been employed. The redox potential of hIAPP was determined as -129 ± 4 mV using the protocol described by Zimmermann *et al.*[16]. (**E**) Western blot of *in vitro* and *in vivo* formed hIAPP aggregates. Left: Purified, recombinant hIAPP after an incubation time of 3 h and 12 h in solution yields SDS-stable hIAPP oligomers. The oligomeric state is increased at prolonged incubation times. Right: Sample extracted from pancreatic islets from diabetic TG/TG and non-diabetic+/+ mice. The oxidation state of cysteines was blocked with NEM in an alkylation reaction which produced a modified $hIAPP^{red}$ exhibiting a larger molecular weight. The high molecular weight bands were observed only in the TG/TG mouse sample. hIAPP was identified using the antibody A133[15]. This antibody is monoclonal against hIAPP(20–29) and is specific for the human sequence. No interference with alkylated hIAPP can thus be expected. (**F**) Light microscopy (left) and histological immunofluorescence images (right) of islets from +/+ control mice and TG/TG mice. Insulin is indicated in green, amyloid fibrils in red, nuclei in blue. +/+ control mice do not display any morphological changes. Amyloid aggregates are observed using an antibody against amyloid fibrils in islets of TG/TG mice. The white scale bar denotes 100 μm. Figure S1 shows that the antibody 91D7E8 specifically detects IAPP fibrils.

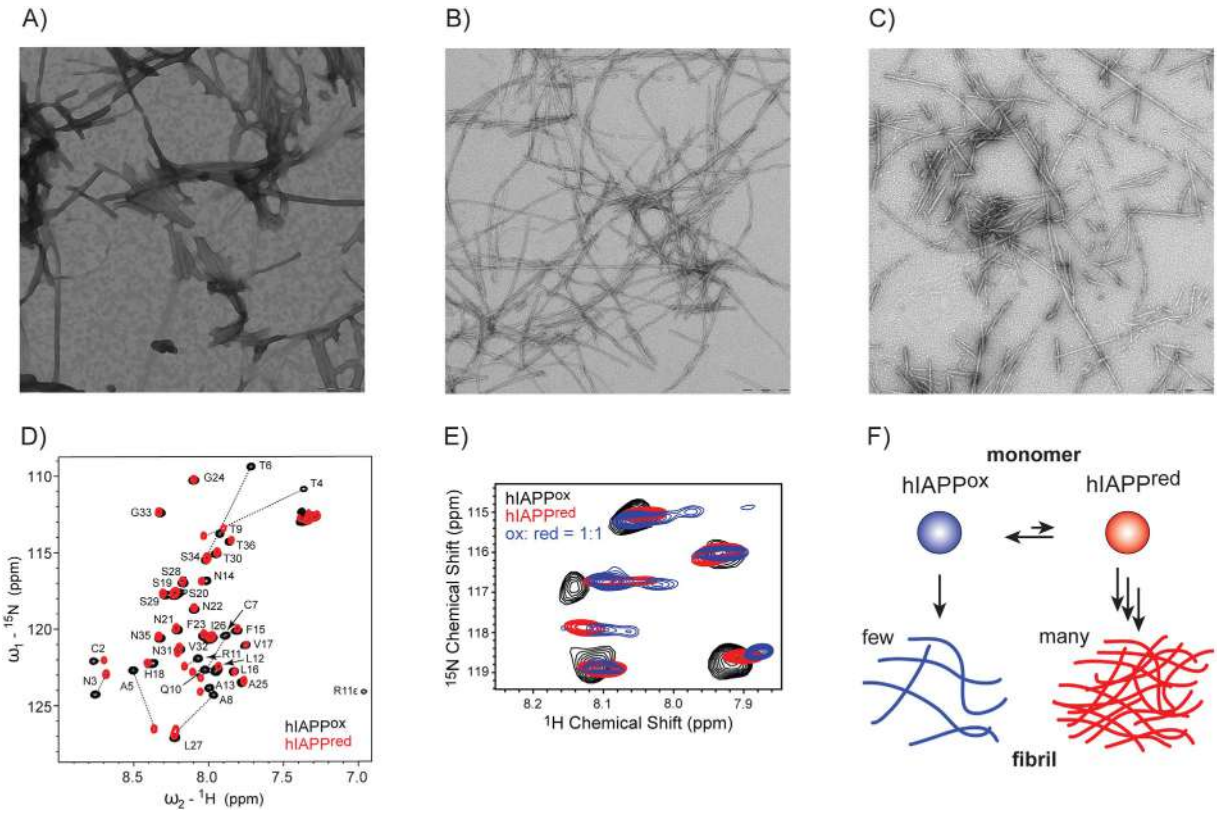


Figure 2. (**A–C**) EM images of fibrils formed in the presence of different ratios of the GSH-GSSG redox couple. In (**A**) hIAPP was incubated using 2 mM GSSG to produce a completely oxidized hIAPP. (**B**) hIAPP was incubated using 35 mM GSH yielding 100% $hIAPP^{red}$, whereas in (**C**) hIAPP was incubated in a buffer containing 0.2 mM GSSG and 4.98 mM GSH yielding 50% $hIAPP^{red}$ and 50% $hIAPP^{ox}$. The scale bar denotes 200 nm. (**D**) $^{1}H,^{15}N$-HSQC spectra of the reduced (red) and oxidized (black) hIAPP in aqueous buffer (30 mM acetic acid, pH 5.3). Amino acids are labelled according to their type and sequence number. Major chemical shift changes upon disulfide bond formation are indicated by black dashed lines. The NMR chemical shifts are deposited in the BMRB (BMRB ID: 34069). (**E**) $^{1}H,^{15}N$ HSQC spectra to probe the hIAPP redox state. Fibrils were prepared under the three redox conditions described in (**A–C**). Before the NMR experiment, fibrils were dissolved in organic solvent. (**F**) Model of the hIAPP aggregation mechanism. ER stress results in production of a small amount of $hIAPP^{red}$ in the cell. $hIAPP^{red}$ aggregates preferentially and is depleted from the pool of soluble molecules in the cell. Redox buffering continuously converts $hIAPP^{ox}$ into $hIAPP^{red}$. This way, hIAPP might assist in regenerating redox equivalents under ER stress conditions.

The same analysis for the reduced form of hIAPP showed a smaller helical propensity for the N-terminal half, suggesting that the N-terminal disulfide bond stabilizes the helical conformation. The observation that $hIAPP^{ox}$ has a diminished tendency to aggregate stimulated a more detailed structural characterization of the peptide. For the structural analysis, 3D $^{13}C/^{15}N$ edited NOESY-HSQC data were collected for $hIAPP^{ox}$. The faster aggregation kinetics of $hIAPP^{red}$ prohibited a detailed structural characterization. The ensemble of the 20 lowest energy structures obtained after water refinement is shown in Fig. 3B,C. A summary of the structural and restraint statistics is given in Table 1. We note that no hydrogen bonding restraints were included in the structure calculations to enforce an α-helical conformation. The present solution structure of native hIAPP (oxidized and amidated) shares similarities, but also has differences when compared to other structures of amylin reported under various conditions[19]. A common theme in all structures is the N-terminal loop closed by the disulfide bond followed by a right-handed α-helix. In our study, the oxidation state of the peptide was confirmed by the $^{13}C\beta$ secondary chemical shifts of the two cysteines (Cys-2 and Cys-7). In our analysis, the α-helix spans residues 8–17. In particular, the helical structure was confirmed by the characteristic αN(i,i+3), αN(i,i+4) and αβ(i,i+3) NOEs (Fig. 3D,G,H). We found that the histidine at position 18 is incompatible with an α-helical conformation. His-18 thus introduces a kink in line with earlier findings (Fig. 3C)[20]. For the C-terminal part, the NOE pattern of the ^{15}N-edited 3D NOESY experiment was dominated by strong correlations between the respective amides and the (i−1) Hα atoms which is typical for IDPs (Intrinsically Disordered Proteins). In addition, no long-range NOEs were detected for the C-terminal amide suggesting that cross-talk to distant residues is not occuring. The structural model is supported by $\{^{1}H\},^{15}N$ heteronuclear NOE data that showed a certain rigidity for the N-terminal half and increased flexibility towards the C-terminus of hIAPP (Fig. 3F).

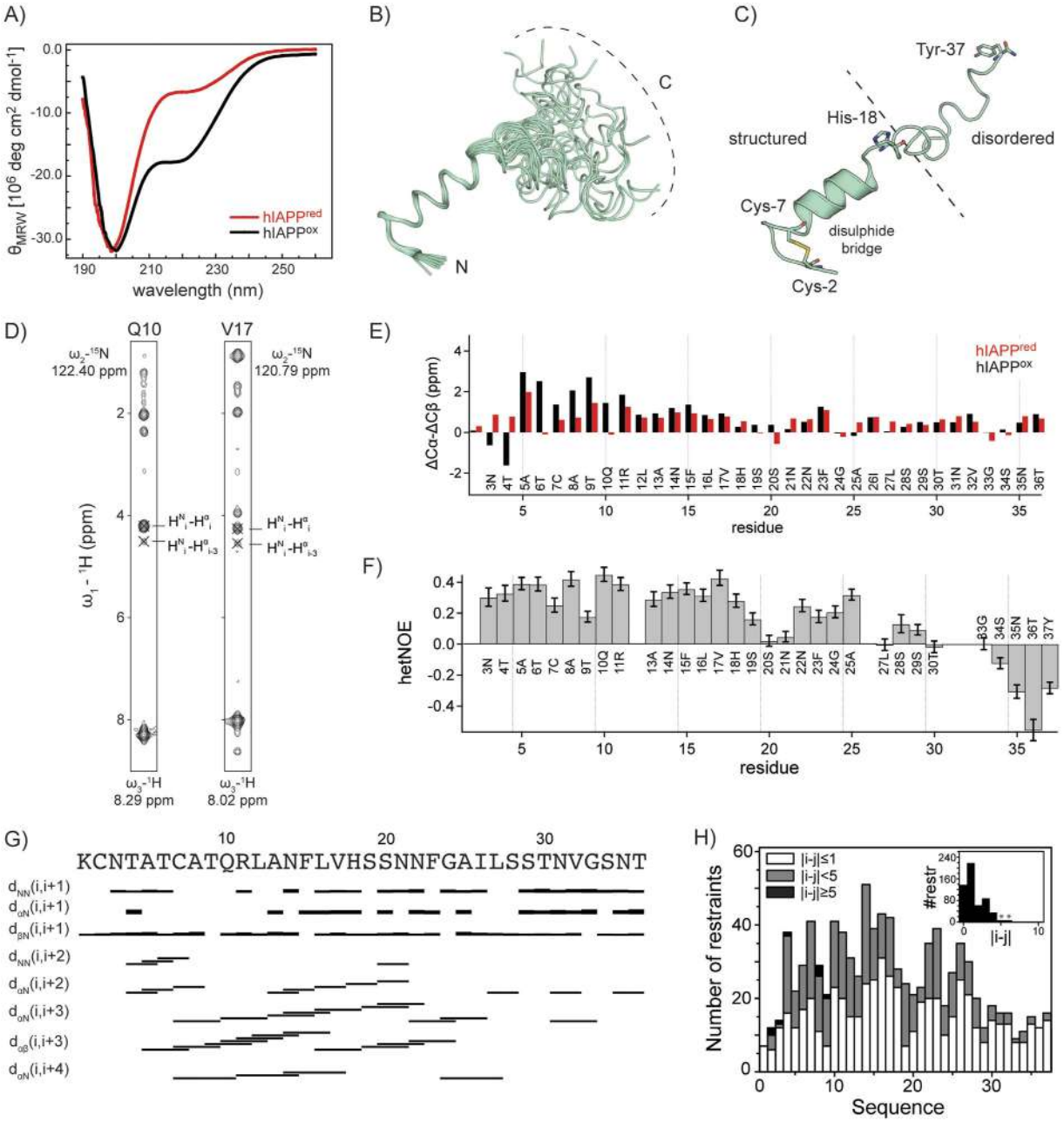


Figure 3. (**A**) Circular dichroism (CD) spectra of reduced (red) and oxidized hIAPP (black). $hIAPP^{ox}$ yields a negative band at 222 nm and 208 nm which is characteristic for an α-helical conformation. The reduced form shows a considerably smaller band at 222 nm indicating a larger degree of random coil conformation. (**B**) NMR structural ensemble of $hIAPP^{ox}$(1-37). The structure was determined at pH 5.3 and at a temperature of 4 °C, using the software package CYANA[21]. (**C**) Structure of one conformer of the hIAPP ensemble. $hIAPP^{ox}$ adopts an α-helix structure involving residues 8–17, and is disordered in its C-terminal part. The disulfide bridge is indicated in yellow. The hIAPP structural ensemble was deposited in the PDB (PDB ID: 5MGQ). RDC experiments confirm that the N-terminal part of hIAPP adopts a compact structure (Fig. S4). (**D**) Representative strips from 3D NOESY experiments. The NOE pattern of the respective residues demarcates the boundaries of the α-helical structure of hIAPP. E) $^{13}C^{\alpha/\beta}$ secondary chemical shift analysis. Differences of deviations from random coil chemical shifts for Cα and Cβ (δΔ, ppm) are represented as a function of the primary sequence. $hIAPP^{red}$ is shown in red. Reduction was induced by adding 10 mM TCEP. The oxidized peptide was obtained by addition of 2 mM H_2O_2 (black). Secondary Structure Propensity (SSP) of $hIAPP^{ox}$ in aqueous buffer, pH 5.3, is shown in Fig. S3. (**F**) Heteronuclear NOEs of hIAPP. The hetNOE data confirms that the N-terminus of hIAPP has a high propensity to form a compact structure. (**G**) Experimental short and medium range NOE connectivities in hIAPP. (**H**) Number of NOE restraints per residue employed for hIAPP structure calculation. The inset displays the number of NOE restraints observed between residue i and j. Only four long range connectivities (residues i and j, with $|i-j|>5$) are observed for the residues around the disulfide bridge (denoted with an asterisk in the inset).

NMR distance and dihedral constraints	
Distance constraints	
Total NOE	542
Intra-residue	136
Inter-residue	406
Sequential ($\|i-j\|=1$)	218
Medium-range ($\|i-j\|<4$)	184
Long-range ($\|i-j\|>5$)	4
Total dihedral angle restraints	20
ϕ	10
ψ	10
Structure statistics	
Violations (mean ± s.d.)	
Distance constraints (Å)	0.019 ± 0.002
Dihedral angle constraints (°)	0.767 ± 0.133
Max. dihedral angle violation (°)	3.02
Max. distance constraint violation (Å)	0.25
Deviations from idealized geometry	
Bond lengths (Å)	0.014 ± 0.001
Bond angles (°)	1.636 ± 0.043
Impropers (°)	1.789 ± 0.155
Average pairwise r.m.s. deviation* (Å)	
Heavy	0.78 ± 0.31
Backbone	0.27 ± 0.09
Ramachandran plot statistics (%)	
Residues in most favoured regions	68.0
Residues in additionally allowed regions	30.6
Residues in generously allowed regions	0.9
Residues in disallowed regions	0.5

Table 1. Structural statistics of hIAPP (residues 1-37). *Pairwise r.m.s. deviation was calculated among 20 refined structures for residues 2–17.

Discussion

Stabilization of the monomeric state of a protein can prevent its conversion into an aggregated state[22]. This paradigm is generally accepted for folded proteins, but has so far not been recognized for disordered peptides and proteins. We show here that the oxidized form of the islet hormone hIAPP is more structured in comparison to the reduced form of the peptide, and at the same time has a smaller propensity to aggregate. We find furthermore that the pancreatic islets of a T2D mouse model contain primarily aggregated hIAPP in its reduced state. We suggest that differential aggregation unleashes redox equivalents that potentially counteract oxidative stress.

Initially, it was suggested that hIAPP adopts a random coil structure[23]. However, more recently reports indicate that the N-terminal part of hIAPP involving residues 5–19 and to a somewhat smaller degree residues 20–22 transiently samples an α-helical structure in solution[8,24]. The same residues adopt a fully helical structure when bound to membranes or micelles[19,24,25]. In agreement with previous studies, we found that reduction of the N-terminal disulfide bond (Cys2–Cys7) in solution decreases the extent of helix formation in this region[26]. Miranker and coworkers observed that formation of helical structure is directly correlated with enhanced amyloid formation on a membrane surface. It was suggested that a parallel association of helices results in locally increased peptide concentration that could nucleate β-strand structure. By contrast, we find that the reduced form of hIAPP aggregates significantly faster than the oxidized peptide. Our data can be reconciled with the observation made by Miranker and coworkers by assuming that membranes can favor an α-helical structure e.g. by formation of an amphipatic helix. At the same time, an increase of the local protein density can be a very strong driving force for aggregation. In aqueous buffer, the situation is fundamentally different. The α-helical propensity at the N-terminus of hIAPP impedes conversion into β-sheet structure. Solid-state NMR studies showed that hIAPP adopts a strand–turn–strand structure in amyloid fibrils, with the N-terminal and C-terminal strands involving residues 8–17 and residues 28–37, respectively[27,28]. The residues in-between the disulfide bridge do not contribute to the amyloid fiber core structure. At the same time, the disulfide bridge plays a central role in the fibril assembly mechanism[29]: Loss of the disulfide bond substantially reduces fiber formation by secondary nucleation, and eliminates the activation step in seeded fiber growth[30].

Obviously, the expression level of hIAPP has a strong impact on solubility and aggregation propensity of the peptide. Insulin and hIAPP tend to be regulated in parallel. Hyperglycemia, however, has been shown to result in disproportionate upregulation of hIAPP versus insulin, thus altering the ratio of hIAPP to insulin secreted by the β-cells[31]. Increased secretion of hIAPP in severe hyperglycemia has been suggested to facilitate the formation of

hIAPP-derived islet amyloid deposits[31]. At the same time, disturbances in redox regulation, calcium homeostasis, glucose deprivation, and viral infection can lead to ER stress[32]. Under these conditions, the folding of proteins is slowed down, leading to an increase of unfolded proteins. Unfolded proteins cause a stress response in the ER which is known as unfolded protein response (UPR)[33]. ER stress is emerging as a potential cause of damage in hypoxia/ischemia, insulin resistance, and other disorders[34]. In T2D, pancreatic β-cells undergo cell death as a consequence of sustained ER stress[35,36]. We find that reduced hIAPP aggregates 3-times faster in comparison to the oxidized form of the peptide. This differential aggregation results in a continuous conversion of $hIAPP^{ox}$ into $hIAPP^{red}$ in the ER. Thus, oxidation equivalents are continuously restored in the ER, a mechanism that might aid to counteract ER stress. Given the fact that hIAPP occurs at concentrations up to 4 mM in the cell[37,38], hIAPP might act as a buffering system. This process works as long as the resulting aggregates are efficiently removed. Clearance can be achieved via autophagy and degradation in the lysosome. It has been shown that β-cell apoptosis is in fact correlated with autophagy and lysosomal degradation[39–41]. Inhibition of autophagy induces apoptosis, whereas stimulation of autophagy protects β-cells from hIAPP-induced toxicity. Alternatively, the endoplasmic reticulum–associated degradation (ERAD) assists to protect cells from ER stress[42]. It has been reported that accumulation of polyubiquitinated proteins occurs in pancreatic islets of hIAPP transgenic mice but not in mice with comparable transgenic expression of rodent-IAPP (r-IAPP). Proteasome dysfunction can mediate obesity induced ER stress and insulin resistance[43]. In fact, it has been found that the β-cell ERAD/ubiquitin/proteasome system of obese individuals with high BMI, that are otherwise healthy and not suffering from T2D, adapts to the increased synthetic burden of insulin and IAPP[44]. p62, an important factor in the autophagy/lysosomal pathway, is downregulated by overproduction of hIAPP in diabetic β-cells, resulting in induction of apoptosis[39,45].

By contrast, the percentage of β-cells with ubiquitin-positive immunostaining was significantly increased in obese individuals with T2D. This may suggest that misregulation of proteasomal degradation yields an accumulation of aggregates which may eventually lead to perturbations in redox balance. *In vivo*, expression of other factors such as BiP, HSP70 and DnaJ which are known to interact with hIAPP can modulate the aggregation behaviour[46]. Similar to the small heat shock proteins such as αB-crystallin that interacts sub-stoichiometrically with Alzheimer's disease Aβ[47], HSP70 and GRP78/BiP can detect and bind misfolded hIAPP oligomers. Chaperones thus would have an influence ensuring proper redox balance.

Recently, it has been shown that the redox status of the microsomal protein folding machinery is affected in diabetes[48]. Streptozotocin-induced diabetes in rats results in a significant increase in the content of free protein sulfhydryl groups, as well as in the amount of free sulfhydryl groups of the major protein disulfide isomerases (PDIs), in particular the 58 kDa PDI and the 57 kDa ERp57. In this chemically induced diabetes model, hIAPP is overexpressed relative to insulin[49]. Our data revealed the redox potential of hIAPP to be on the order of (-129 ± 4) mV suggesting that a significant amount of hIAPP in the ER actually exists in the reduced state. A shift of the redox conditions under ER stress to a more reducing state, and an increase in the amount of produced hIAPP in T2D would immediately result in enhanced hIAPP aggregation. As a consequence of aggregation, even more redox equivalents would become available to restore the oxidizing redox environment of the ER.

Taken together, we have shows that aggregation of hIAPP can have beneficial effects on the well-being of the cell by restoring redox equivalents in the due course of aggregation. Only failure of the preoteasome degradation machinery results in β-cell disruption. It remains to be seen which molecular mechanisms in more detail induce a misfunction of proteasomal degradation in T2D.

Materials and Methods

Recombinant protein expression and purification. Human IAPP was expressed in *E. coli* and purified using the protocol described previously[17]. This protocol allows the production of an isotopically enriched peptide, which is amidated at the C-terminus and is disulfide bridged involving residues Cys-2 and Cys-7. Molecular biology reagents were obtained from Roche, New England Biolabs and from Sigma-Aldrich St. Louis, MO, USA. Isotopically labeled minimal media components were purchased from Cambridge Isotope Laboratories (CIL).

Isolation of mouse islets. hIAPP transgenic mice were purchased from Jackson Laboratory (Bar Harbor, USA) and studied between 13–18 weeks of age. Animals were killed by isoflurane overdose and islets were isolated immediately. To perfuse the pancreas, the common bile duct was clamped at the papilla of Vater and 1 mg/ml cold collagenase P solution in HBSS solution supplemented with 1% BSA was injected into the common bile duct. The pancreas was removed and incubated for 15 min at 37 °C. Digestion was stopped by addition of a cold HBSS solution supplemented with 1% BSA and centrifuged at 290 g for 2 min at 4 °C. The pellet was resuspended in the HBSS solution supplemented with 1% BSA, filtered through a metal mesh and centrifuged at 290 g for 2 min at 4 °C. The pellet was resuspended in 15% Optiprep supplemented with 3.75 mM HEPES, 12% PBS, 6.2% RPMI and layered onto 15% Optiprep supplemented with 3.75 mM HEPES, 12% PBS, 6.2% RPMI solution. The pancreatic lysate was overlaid with the HBSS solution supplemented with 1% BSA, incubated for 10 min and centrifuged at 360 g for 18 min at 4 °C. After centrifugation, islets were observed in the interface between the first and second layers and after collecting islet fraction, cell debris were filtered through a 70 μm cell strainer and islets were washed from the filter with RPMI supplemented with 11 mM glucose, 10% FBS, antibiotic-antimycotic solution. To eliminate remaining exocrine tissue, the islets were handpicked and transferred into a new culture dish.

Animals were declared to the responsible local authority (Regierung von Oberbayern). The Institutional Animal Welfare Officer (Helmholtz-Zentrum München) was consulted about the work and approved the study. All experiments were performed in accordance with relevant guidelines and regulations.

Western Blot Analysis. Extracted islets were solubilized with 30 μL PBS buffer supplemented with NEM and vortexed. Subsequently, 10 μL of SDS sample buffer without reducing agent was added. Samples were boiled for 5 min at 90 °C and separated by 12% Mini-PROTEAN-TGX stain-Free™ protein gels, and then transferred

onto a Tnas-Blot-Turbo™ Mini PVDF membrane with a Bio-Rad Trans-Blot Turbo™ transfer system. Membrane was blocked in Tween-Tris-buffered saline with 5% (w/v) nonfat dry milk (Roth) for 1 hour and then incubated with the primary antibody A133 in Tween-Tris-buffered saline with a ratio 1:1000 overnight at 4 °C. Reactivity was analyzed with antirabbit peroxidase-conjugated secondary antibody (Serva) and chemiluminescence detection (Bio-Rad).

hIAPP fibril antibody production. To generate a monoclonal antibody against hIAPP amyloid fibrils, synthetic human lyophilized HFIP-treated hIAPP was dissolved in 20 mM Tris buffer (pH 7.4) to a concentration of 30 μM. The solution was sonicated for 1 min and incubated for 2 d at 37 °C with shaking at 300 rpm. The resulting fibrils were used for antibody generation according to a proprietary immunization protocol of Synaptic Systems Göttingen (see also http://www.sysy.com/mabservice.html): Three 8–10-week-old BALB/c female mice were subcutaneously immunized with IAPP fibrils over a period of 17 days. Cells from the knee lymph nodes were fused with the mouse myeloma cell line P3 × 63Ag8.653 (ATCC CRL-1580). The clones used in this study were re-cloned two times by limiting dilution and the immunoglobulin subclass was determined. Antibodies secreted by the hybridomas were screened for their reactivity against the immunogen using ELISA. Positive antibodies were retested by ELISA against hIAPP fibrils and monomer. Hybridomas producing antibodies with preference for IAPP fibrils were subcloned to monoclonality and further analyzed for histology and western blot applications. One of the antibodies showing high specificty for hIAPP fibrils was the antibody 91D7E8.

Immunofluorescence staining of pancreatic β-cells. Pancreata were fixed in 4% paraformaldehyde and processed as published previously[50]. Briefly, cryo-sections were stained with anti-insulin (Invitrogen, Life Technologies, Germany), or with anti-amyloid fibril antibody 91D7E8 employing corresponding fluorescent-labeled secondary antibodies (Invitrogen, Life Technologies, Germany). DAPI was used to visualize cell nuclei. Slides were scanned using a NanoZoomer 2.0 HT (Hamamatsu, Japan) fluorescence scanner.

NMR sample preparation. Lyophilized hIAPP(1-37) was dissolved into NMR buffer (30 mM deuterated acetic acid, pH 5.3) containing 10% D_2O for locking. The samples were measured and stored at 4 °C. The formation of the intramolecular disulfide bond was confirmed by NMR. To prepare hIAPP in the oxidized state, a hIAPP solution was supplemented with 0.2 mM GSSG. To generate the reduced form of the peptide, the GSH powder was directly added to the hIAPP solution to yield a final concentration of 35 mM of reduced glutathione. To produce the $hIAPP^{red}$:$hIAPP^{ox}$ = 50:50 mixed sample, the hIAPP stock solution was added to a glutathione solution with final concentrations of 0.2 mM GSSG and 4.98 mM GSH. The final hIAPP concentration amounted to 100 μM in all cases. Afterwards, the samples were incubated at 4 °C to allow for complete oxidation or reduction.

For analysis of the oxidation state of the aggregated hIAPP samples at varying redox conditions, peptides were incubated first for two weeks at 37 °C to allow for fibril formation. After this period, samples were centrifuged for 30 min to harvest the fibrils. Pellets were dissolved in 250 μL of d_6-DMSO, and subsequently sonicated for 1 hour. In order to prevent degradation of hIAPP, ice was added into the sonication bath to avoid sample heating. Afterwards, samples were directly transferred into Shigemi NMR tubes (Shigemi Inc., Allison Park, USA) for solution-state NMR measurements.

NMR Experiments. All NMR experiments were performed at 4 °C, employing Bruker Avance 500, 600, 750 MHz spectrometers, equipped with cryo-probes. The proton chemical shifts were referenced to the water resonance frequency, the ^{15}N and ^{13}C shifts were referenced indirectly. Backbone and side chain assignments were obtained using triple resonance experiments[51]. Side chain assignments were obtained from 3D-HCCH TOCSY experiments. Overall, an assignment completeness of 97% was obtained. NMR spectra were processed using the software TopSpin (Bruker) and NMRPIPE[52]. Spectra were analyzed using SPARKY[53] and ccpNMR analysis[54].

Structure Calculation. Sequential and inter-residual NOE cross-peaks that could be resolved in the ^{15}N-edited spectrum were assigned manually for the N-terminal half and were kept fixed during structure calculations with CYANA[21]. Peak intensities for all cross-peaks were converted to distances using the internal function of CYANA. Dihedral angle restraints were derived from chemical shifts using the program TALOS[55,56]. Distance and angle restraints were used as input for the calculation of the structure on a linux-based cluster. All non-violated structures were recalculated using a water-refinement protocol.

Transmission electron microscopy (TEM). Samples were stained with 1% uranyl acetate solution on Formvar/Carbon 300 mesh copper coated grids from Electron Microscopy Sciences (Hatfield, PA 19440, USA). 10 μL of the sample were placed on the grid for 1 min. The excess of the sample was dried using filter paper. After this process, the grid was washed by placing it three times on a drop of water for 3 s. For staining, the grid was placed on a drop of a uranyl acetate solution for 30 s. The excess of the solution was again dried using filter paper. Samples were measured immediately employing an EM 10 CR (Zeiss, Germany).

Determination of the hIAPP redox potential. 100 μM $hIAPP^{ox}$ was initially dissolved in 500 μL NMR buffer, containing 0.2 mM oxidized glutathione (GSSG). The GSH concentration was then increased stepwise to yield a final GSH concentration of 0.2, 1.0, 3.0, 5.0, 7.0, 10.0, 12.0, 15.0 and 18.0 mM, using a 100 mM GSH stock solution. ^{1}H,^{15}N-HSQC spectra were recorded after each addition of GSH. To determine the redox potential, the intensity of Cys-7 was determined in each spectrum. The intensities of the Cys-7 cross peak were plotted as a function of $[GSH]^2/[GSSG]$ on a logarithmic scale, and fit assuming a sigmoidal behavior. The $[GSH]^2/[GSSG]$ titration half-point was determined as 0.24 M. The redox potential for the GSH/GSSG redox pair was assumed to be −151 mV[13]. The Nernst equation yields then a redox potential for hIAPP of −129+/−4 mV.

Circular Dichroism Spectroscopy (CD). CD experiments were carried out at room temperature employing a JASCO spectropolarimeter. Quartz cells with a path length of 0.1 cm were used. Spectra were recorded between 260 nm and 190 nm at 0.1 nm intervals, with a response time of 1 s. For the analysis of the data the background buffer was subtracted and the data were expressed as mean residue ellipticity $\theta_{MRW} = (\text{deg cm}^2\,\text{dmol}^{-1}) \times 10^{-6}$.

Thioflavin-T assay. Fluorescence experiments were performed using a Beckman fluorometer, using Quartz cells with a path length of 1 cm. The samples were freshly prepared with the dye ThT, and incubated in the cell while stirring at 37 °C. The solutions were prepared by adding NMR buffer (30 mM d-acetic acid, pH 5.3) to the dried peptide immediately prior to the measurements, holding the solution on ice all the time. A ThT stock solution was prepared in NMR buffer at a concentration of 25 mM. Final concentrations for the ThT experiment were 30 μM hIAPP and 10 μM ThT. The data were analyzed using the software Origin (OriginLab Corporation, Northampton, MA 01060, USA), and were fit assuming a sigmoidal behavior.

References

1. Amos, A. F., McCarty, D. J. & Zimmet, P. The rising global burden of diabetes and its complications: estimates and projections to the year 2010. *Diabetic Med.* **14** Suppl 5, S1–S85 (1997).
2. Donath, M. Y. & Shoelson, S. E. Type 2 diabetes as an inflammatory disease. *Nature Rev. Immunol.* **11,** 98–107 (2011).
3. Paulsson, J. F., Andersson, A., Westermark, P. & Westermark, G. T. Intracellular amyloid-like deposits contain unprocessed pro-islet amyloid polypeptide (proIAPP) in beta cells of transgenic mice overexpressing the gene for human IAPP and transplanted human islets. *Diabetologia* **49,** 1237–1246 (2006).
4. Cooper, G. J. S. *et al.* Purification and Characterization of a Peptide from Amyloid-rich Pancreases of Type-2-Diabetic Patients. *Proc. Natl Acad. Sci. USA* **84,** 8628–8632 (1987).
5. Bram, Y. *et al.* Apoptosis induced by islet amyloid polypeptide soluble oligomers is neutralized by diabetes-associated specific antibodies. *Sci. Rep*. **4** (2014).
6. Hutton, J. C. The Insulin Secretory Granule. *Diabetologia* **32,** 271–281 (1989).
7. Westermark, P., Li, Z. C., Westermark, G. T., Leckstrom, A. & Steiner, D. F. Effects of beta cell granule components on human islet amyloid polypeptide fibril formation. *FEBS Lett*. **379,** 203–206 (1996).
8. Wiltzius, J. J. W., Sievers, S. A., Sawaya, M. R. & Eisenberg, D. Atomic structures of IAPP (amylin) fusions suggest a mechanism for fibrillation and the role of insulin in the process. *Prot. Sci.* **18,** 1521–1530 (2009).
9. Jha, S. *et al.* pH Dependence of Amylin Fibrillization. *Biochemistry* **53,** 300–310 (2014).
10. Braakman, I., Helenius, J. & Helenius, A. Manipulating Disulphide Bond Formation and Protein Folding in the Endoplasmic Reticulum. *EMBO J.* **11,** 1717–1722 (1992).
11. Cuozzo, J. W. & Kaiser, C. A. Competition between glutathione and protein thiols for disulphide-bond formation. *Nature Cell Biol*. **1,** 130–135 (1999).
12. Nardai, G., Korcsmaros, T., Papp, E. & Csermely, P. Reduction of the endoplasmic reticulum accompanies the oxidative damage of diabetes mellitus (Reprinted from Thiol Metabolism and Redox Regulation of Cellular Functions). *Biofactors* **17,** 259–267 (2003).
13. Schafer, F. Q. & Buettner, G. R. Redox environment of the cell as viewed through the redox state of the glutathione disulfide/glutathione couple. *Free Radic. Biol. Med.* **30,** 1191–1212 (2001).
14. Birk, J. *et al.* Endoplasmic reticulum: reduced and oxidized glutathione revisited. *J. Cell Sci.* **126,** 1604–1617 (2013).
15. Paulsson, J. F. & Westermark, G. T. Aberrant processing of human proislet amyloid polypeptide results in increased amyloid formation. *Diabetes* **54,** 2117–2125 (2005).
16. Zimmermann, J., Kuhne, R., Sylvester, M. & Freund, C. Redox-regulated conformational changes in an SH3 domain. *Biochemistry* **46,** 6971–6977 (2007).
17. Rodriguez Camargo, D. C. *et al.* Cloning, expression and purification of the human Islet Amyloid Polypeptide (hIAPP) from Escherichia coli. *Protein Expr. Purif.* **106,** 49–56 (2015).
18. Grzesiek, S. & Bax, A. An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. Journal of Magnetic Resonance *J. Magn. Reson.* **99,** 201–207 (1992).
19. Nanga, R. P. R., Brender, J. R., Vivekanandan, S. & Ramamoorthy, A. Structure and membrane orientation of IAPP in its natively amidated form at physiological pH in a membrane environment. *Biochim. Biophys. Acta* **1808,** 2337–2342 (2011).
20. Abedini, A. & Raleigh, D. P. The role of His-18 in amyloid formation by human islet amyloid polypeptide. *Biochemistry* **44,** 16284–16291 (2005).
21. Guntert, P. Automated NMR structure calculation with CYANA. *Meth. Mol. Biol.* **278,** 353–78 (2004).
22. Chiti, F. & Dobson, C. M. Amyloid formation by globular proteins under native conditions. *Nature Chemical Biology* **5,** 15–22 (2009).
23. Kayed, R. *et al.* Conformational transitions of islet amyloid polypeptide (IAPP) in amyloid formation *in vitro*. *J. Mol. Biol.* **287,** 781–796 (1999).
24. Williamson, J. A. & Miranker, A. D. Direct detection of transient a-helical states in islet amyloid polypeptide. *Protein Science* **16,** 110–117 (2007).
25. Patil, S. M., Xu, S. H., Sheftic, S. R. & Alexandrescu, A. T. Dynamic alpha-Helix Structure of Micelle-bound Human Amylin. *J. Biol. Chem.* **284,** 11982–11991 (2009).
26. Williamson, J. A., Loria, J. P. & Miranker, A. D. Helix Stabilization Precedes Aqueous and Bilayer-Catalyzed Fiber Formation in Islet Amyloid Polypeptide. *J. Mol. Biol.* **393,** 383–396 (2009).
27. Luca, S., Yau, W.-M., Leapman, R. D. & Tycko, R. Peptide Conformation and Supramolecular Organization in Amylin Fibrils: Constraints from Solid-State NMR. *Biochemistry* **46,** 13505–13522 (2007).
28. Weirich, F. *et al.* Structural Characterization of Fibrils from Recombinant Human Islet Amyloid Polypeptide by Solid-State NMR: The Central FGAILS Segment Is Part of the β-Sheet Core. *PLoS One* **11,** e0161243 (2016).
29. Koo, B. W. & Miranker, A. D. Contribution of the intrinsic disulfide to the assembly mechanism of islet amyloid. *Prot. Sci.* **14,** 231–239 (2005).
30. Padrick, S. B. & Miranker, A. D. Islet amyloid: Phase partitioning and secondary nucleation are central to the mechanism of fibrillogenesis. *Biochemistry* **41,** 4694–4703 (2002).
31. Obrien, T. D., Westermark, P. & Johnson, K. H. Islet Amyloid Polypeptide and Insulin-Secretion from Isolated Perfused Pancreas of Fed, Fasted, Glucose-treated and Dexamethasone-treated Rats. *Diabetes* **40,** 1701–1706 (1991).
32. Xu, C. Y., Bailly-Maitre, B. & Reed, J. C. Endoplasmic reticulum stress: cell life and death decisions. *J. Clin. Invest.* **115,** 2656–2664 (2005).
33. Walter, P. & Ron, D. The Unfolded Protein Response: From Stress Pathway to Homeostatic Regulation. *Science* **334,** 1081–1086 (2011).
34. Ozcan, U. *et al.* Endoplasmic reticulum stress links obesity, insulin action, and type 2 diabetes. *Science* **306,** 457–461 (2004).
35. Butler, A. E. *et al.* Beta-cell deficit and increased beta-cell apoptosis in humans with type 2 diabetes. *Diabetes* **52,** 102–110 (2003).

36. Westermark, P., Andersson, A. & Westermark, G. T. Islet Amyloid Polypeptide, Islet Amyloid and Diabetes Mellitus. *Physiol. Rev.* **91,** 795–826 (2011).
37. Hutton, J. C. Secretory Granules. *Experientia* **40,** 1091–1098 (1984).
38. Clark, A. & Moffitt, J. Pancreatic Islet Amyloid and Diabetes. In *Protein Misfolding, Aggregation, and Conformational Diseases. Part B: Molecular Mechanisms of Conformational Diseases*, vol. 6 (eds Uversky, V. N. & Fink, A. L.) 199–216 (Springer, New York, 2007).
39. Rivera, J. F. *et al.* Human-IAPP disrupts the autophagy/lysosomal pathway in pancreatic beta-cells: protective role of p62-positive cytoplasmic inclusions. *Cell Death Diff.* **18,** 415–426 (2011).
40. Morita, S. *et al.* Autophagy protects against human islet amyloid polypeptide-associated apoptosis. *J. Diabetes Investig.* **2,** 48–55 (2011).
41. Lee, M. S. Role of islet beta cell autophagy in the pathogenesis of diabetes. *Trends Endocrinol. Metabol.* **25,** 620–627 (2014).
42. Huang, C. J. *et al.* Induction of endoplasmic reticulum stress-induced beta-cell apoptosis and accumulation of polyubiquitinated proteins by human islet amyloid polypeptide. *Am. J. Physiol. Endocrinol. Metab.* **293,** E1656–E1662 (2007).
43. Otoda, T. *et al.* Proteasome Dysfunction Mediates Obesity-Induced Endoplasmic Reticulum Stress and Insulin Resistance in the Liver. *Diabetes* **62,** 811–824 (2013).
44. Costes, S. *et al.* Beta-Cell Dysfunctional ERAD/Ubiquitin/Proteasome System in Type 2 Diabetes Mediated by Islet Amyloid Polypeptide-Induced UCH-L1 Deficiency. *Diabetes* **60,** 227–238 (2011).
45. Rivera, J. F., Costes, S., Gurlo, T., Glabe, C. G. & Butler, P. C. Autophagy defends pancreatic beta cells from human islet amyloid polypeptide-induced toxicity. *J. Clin. Invest.* **124,** 3489–3500 (2014).
46. Chien, V. *et al.* The chaperone proteins HSP70, HSP40/DnaJ and GRP78/BiP suppress misfolding and formation of beta-sheet-containing aggregates by human amylin: a potential role for defective chaperone biology in Type 2 diabetes. *Biochem. J.* **432,** 113–121 (2010).
47. Mainz, A. *et al.* The Chaperone αB-Crystallin Deploys Different Interfaces to Capture an Amorphous and an Amyloid Client. *Nature Struct. Mol. Biol.* **22,** 898–905 (2015).
48. Nardai, G. *et al.* Diabetic changes in the redox status of the microsomal protein folding machinery. *Biochem. Biophys. Res. Comm.* **334,** 787–795 (2005).
49. Mulder, H., Ahren, B. & Sundler, F. Islet amyloid polypeptide (amylin) and insulin are differentially expressed in chronic diabetes induced by streptozotocin in rats. *Diabetologia* **39,** 649–657 (1996).
50. Franko, A. *et al.* Bezafibrate improves insulin sensitivity and metabolic flexibility in STZ-treated diabetic mice. *Diabetes* **65,** 2540–2552 (2016).
51. Sattler, M., Schleucher, J. & Griesinger, C. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. NMR Spect.* **34,** 93–158 (1999).
52. Goddard, T. D. & Kneller, D. G. *SPARKY 3*. (University of California, San Francisco, 1997).
53. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6,** 277–293 (1995).
54. Vranken, W. F. *et al.* The CCPN Data Model for NMR Spectroscopy: Development of a Software Pipeline. *Proteins* **59,** 687–696 (2005).
55. Cornilescu, G., Delaglio, F. & Bax, A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13,** 289–302 (1999).
56. Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR* **44,** 213–223 (2009).

Acknowledgements

We are grateful to Gunilla Westermark for providing us with a sample of the antibody A133. Furthermore, we would like to thank Dr. Vanessa Morris and Ashish Kawale for stimulating discussions. This work was performed in the framework of the SFB-1035 (Project-B07, Deutsche Forschungsgemeinschaft, DFG, to B.R.). We acknowledge support of the Helmholtz-Gemeinschaft, and the DFG (Grants Re1435, to B.R.). We are grateful to the Center for Integrated Protein Science Munich (CIPS-M) for financial support. K.T. acknowledges support from the Ministry of Education, Youth and Sports of the Czech Republic under the project CEITEC 2020 (LQ1601). This work was supported by the Bavarian Ministry of Sciences, Research and the Arts (Bavarian Molecular Biosystems Research Network, to T.M.), the DFG (Emmy Noether program MA 5703/1-1 to T.M. and CO 291/5-1 to M.C.), and the Austrian Science Fund (FWF: P28854). C.G. gratefully acknowledges the postdoctoral fellowship program (PFP) of the Helmholtz-Zentrum-München. We thank the HELENA and the TUM graduate schools for support.

Author Contributions

D.C.R.C. cloned, recombinantly expressed, purified and characterized hIAPP. D.C.R.C., K.T., R.S., and C.H. performed NMR experiments and analyzed the data. K.T., C.H. and C.G. carried out structure calculations. A.F., M.S. and M.H.d.A. contributed fluorescence and light microscopy experiments using islets from hIAPP transgenic mice. D.C.R.C., K.B. and M.C. performed the redox state analysis of *in vivo* material. A.B., C.E., H.M., and E.E.W. produced the hIAPP fibril selective antibody. D.C.R.C., M.A., G.M. and A.K.W. carried out EM experiments. D.C.R.C., K.T., T.M., M.C., E.E.W., M.H.d.A., and B.R. designed the experiments. D.C.R.C. and B.R. wrote the manuscript with contributions from all authors.

Additional Information

Accession codes: The atomic coordinates and restraint files of the oxidized and amidated hIAPP are deposited in the PDB with accession code 5MGQ, chemical shifts at BMRB, entry 34069.

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing Interests: The authors declare no competing financial interests.

How to cite this article: Camargo, D. C. R. *et al.* The redox environment triggers conformational changes and aggregation of hIAPP in Type II Diabetes. *Sci. Rep.* **7**, 44041; doi: 10.1038/srep44041 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.


This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2017

nature COMMUNICATIONS

ARTICLE

Received 9 Apr 2016 | Accepted 19 Jan 2017 | Published 10 Mar 2017

DOI: 10.1038/ncomms14635 OPEN

Allosteric modulation of peroxisomal membrane protein recognition by farnesylation of the peroxisomal import receptor PEX19

Leonidas Emmanouilidis[1,2,*], Ulrike Schütz[1,2,*], Konstantinos Tripsianes[3,*], Tobias Madl[1,2,4], Juliane Radke[5], Robert Rucktäschel[5], Matthias Wilmanns[6], Wolfgang Schliebs[5], Ralf Erdmann[5] & Michael Sattler[1,2]

The transport of peroxisomal membrane proteins (PMPs) requires the soluble PEX19 protein as chaperone and import receptor. Recognition of cargo PMPs by the C-terminal domain (CTD) of PEX19 is required for peroxisome biogenesis *in vivo*. Farnesylation at a C-terminal CaaX motif in PEX19 enhances the PMP interaction, but the underlying molecular mechanisms are unknown. Here, we report the NMR-derived structure of the farnesylated human PEX19 CTD, which reveals that the farnesyl moiety is buried in an internal hydrophobic cavity. This induces substantial conformational changes that allosterically reshape the PEX19 surface to form two hydrophobic pockets for the recognition of conserved aromatic/aliphatic side chains in PMPs. Mutations of PEX19 residues that either mediate farnesyl contacts or are directly involved in PMP recognition abolish cargo binding and cannot complement a Δ*PEX19* phenotype in human Zellweger patient fibroblasts. Our results demonstrate an allosteric mechanism for the modulation of protein function by farnesylation.

[1] Institute of Structural Biology, Helmholtz Zentrum München, Ingolstädter Landstr. 1, Neuherberg 85764, Germany. [2] Munich Center for Integrated Protein Science at Chair of Biomolecular NMR, Department Chemie, Technische Universität München, Lichtenbergstr. 4, Garching 85747, Germany. [3] CEITEC—Central European Institute of Technology, Masaryk University, Kamenice 5, Brno 62500, Czech Republic. [4] Institute of Molecular Biology and Biochemisty, Medical University of Graz, Graz 8010, Austria. [5] Institute of Biochemistry and Pathobiochemistry, Department of Systems Biology, Faculty of Medicine, Ruhr University Bochum, Bochum 44780, Germany. [6] EMBL Hamburg, Notkestr. 85, Geb. 25A, 22607 Hamburg, Germany. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to K.T. (email: kostas.tripsianes@ceitec.muni.cz) or to M.S. (email: sattler@helmholtz-muenchen.de).

Peroxisomes are ubiquitous organelles present in all eukaryotic cells with pivotal roles in cellular homoeostasis. They catalyse reactions in lipid metabolism and decompose hydrogen peroxide, as well as numerous other toxic compounds. The biological importance of peroxisomes is highlighted by a number of inherited diseases associated with malfunctions of peroxisomal proteins[1]. Single-enzyme defects like acatalasia can give rise to comparably mild phenotypes, whereas defects in peroxisomal biogenesis, for example, in the Zellweger syndrome, are lethal. Peroxisomal proteins are directed to the organelle posttranslationally via distinct transport systems, which are specific for either matrix or membrane proteins[2]. The molecular mechanisms of matrix protein recognition for transport into peroxisomes are well characterized. Cycling receptors recognize the matrix proteins in the cytosol, direct them to a docking complex at the peroxisomal membrane, where import of the folded proteins takes places through a transient import pore[3,4]. In comparison, our knowledge on the transport of peroxisomal membrane proteins (PMPs) is still scarce. Peroxisomal biogenesis factor 19 (PEX19) undoubtedly is a key player in several steps of PMP transport. First, it is thought to function as a chaperone for newly synthesized PMPs in the cytosol[5]. Second, PEX19 directs the cargo to the peroxisomal membrane, where it docks to the transmembrane protein PEX3 thereby acting as a shuttling receptor[6]. Third, it could be involved in membrane insertion of PMPs (refs 7,8). Finally, the transfer of PEX3 from the ER to the peroxisome, which is an early step in peroxisome biogenesis, occurs in a PEX19-dependent manner[9].

PEX19 is posttranslationally modified by farnesylation. In spite of an overall low sequence similarity across PEX19 homologues, the farnesylation site is conserved throughout evolution with an exception of trypanosomal PEX19 (ref. 10). The farnesyl group is a C15 isoprenoid (Fig. 1a), which is covalently attached to target proteins by farnesyltransferase[11]. This enzyme catalyses the attachment of the farnesyl group from farnesyl pyrophosphate (substrate) to a cysteine residue of a C-terminal signal sequence called CaaX box ('C' denotes the modified Cys, 'a' an aliphatic amino acid and 'X' usually stands for Ser, Thr, Gln, Ala or Met). Farnesylation and geranylgeranylation, a C20 isoprenoid modification, are classified as prenylation and are irreversible post-translational modifications[12] mostly found in small GTPases, that is, Ras and Rho proteins[13]. Human PEX19 is farnesylated *in vivo*[14,15]. It was reported that PEX19 is completely farnesylated in *Saccharomyces cerevisiae* and that the affinity for PMP cargo peptides is ten-fold increased with farnesylated PEX19 (ref. 16). Furthermore, farnesylation-deficient yeast exhibits reduced stability of PMPs and suffers from defects in peroxisomal biogenesis *in vivo*[16]. It has been suggested that farnesylation may alter the conformation of PEX19, but the molecular and functional consequences of PEX19 farnesylation are so far unknown.

Human PEX19 consists of 299 amino acids with an intrinsically disordered N-terminal half that interacts with the membrane-bound docking protein PEX3 (refs 6,17,18) and PEX14 (ref. 19). The folded C-terminal domain (CTD) mediates binding to PMPs and harbours the CaaX box site for farnesylation[20] (Fig. 1a). A crystal structure of a C-terminal fragment (comprising residues 161–283)[21], referred to as CTDΔC below, showed that the PEX19 fold comprises four α-helices that exhibit a three-helical bundle domain and an additional N-terminal helix α1 protruding away from this bundle. *In vitro* binding studies demonstrated that the CTDΔC fragment is capable of binding PMP peptides with micromolar affinity, while mutational analysis suggested that residues located in helix α1 contribute to cargo binding[21]. However, the crystallized PEX19 fragment lacks the C-terminal 16 amino acids, which include the farnesylation site, and thus the structural impact of farnesylation and its role in the modulation of PMP binding remain unknown.

Here, we present the solution structure of the farnesylated C-terminal PMP binding domain of human PEX19 and report molecular details for the recognition of hydrophobic residues in PMPs determined by NMR spectroscopy. NMR data indicate that the C-terminal residues of the CTD become rigid upon farnesylation. Surprisingly, the PEX19 CTD undergoes significant conformational changes to accommodate the farnesyl group inside a large hydrophobic cavity, which, in turn, affect functional interactions of PEX19 with PMPs. On the basis of NMR chemical shift perturbations and intermolecular NOEs we identify two hydrophobic binding pockets for aromatic residues in PMPs, which are formed upon farnesylation of PEX19. Mutations interfering with either farnesyl recognition or PMP interactions affect the PMP binding *in vitro* and the biological activity of PEX19 during peroxisome biogenesis *in vivo*. Our data suggest that farnesylation orchestrates the PMP binding region of PEX19 by an allosteric mechanism. Thereby, farnesylation contributes PMP binding and the functional activity of PEX19 in PMP import. These findings reveal a novel mechanism for the modulation of protein function by farnesylation.

Results

Structure of the farnesylated PEX19 CTD. To investigate the effects of farnesylation, we studied the human PEX19 CTD (residues 161–299) that includes the native C-terminus with the CaaX box of the protein (Fig. 1a). PEX19 CTD is structurally autonomous and does not interact with the preceding disordered region of the full-length protein as judged by NMR fingerprint spectra (Supplementary Fig. 1a). Recombinant farnesyltransferase was used for farnesylation of PEX19 CTD *in vitro*. An increased electrophoretic mobility in SDS–polyacrylamide gel electrophoresis (SDS–PAGE) (Supplementary Fig. 1b) and NMR spectra (Fig. 1b) demonstrate virtually complete farnesylation of the PEX19 CTD. NMR $^{13}C^{\alpha/\beta}$ secondary chemical shifts indicate that the four α-helices observed in the crystal structure of the truncated PEX19 CTDΔC (residues 161–283) are also present in the farnesylated protein (Supplementary Fig. 1c,d). However, farnesylation induces large differences in amide chemical shifts throughout the PEX19 CTD that involve many residues remote from the farnesylation site (Cys296) (Fig. 1b,c). Low $\{^1H\}$-^{15}N heteronuclear NOE values and large solvent paramagnetic relaxation enhancements (solvent PREs) for the C-terminal ten residues in the non-farnesylated PEX19 CTD indicate that they are highly flexible in solution and solvent exposed, respectively. In contrast, a significant increase of heteronuclear NOE, and reduced sPRE values in the farnesylated PEX19 CTD (Fig. 1c) demonstrate that farnesylation strongly reduces the backbone flexibility and solvent accessibility of the C-terminal residues. Similar sPRE changes, although to a lesser extent, are also observed for residues in the N-terminal helix α1 (Fig. 1c). These data demonstrate that farnesylation induces a substantial rigidification and compaction of the PEX19 CTD that primarily stabilizes the C-terminal region but also locks the arrangement of helix α1 with respect to the core domain. Comparable tumbling correlation times (τ_c) derived from ^{15}N NMR relaxation data for non-farnesylated ($\tau_c = 11.4$ ns) and farnesylated ($\tau_c = 10.5$ ns) PEX19 CTD demonstrate that both protein conformations are monomeric in solution (Supplementary Fig. 1c,d).

To unravel the structural changes induced by farnesylation we determined the structure of the farnesylated PEX19 CTD (Fig. 2a–c). Given the presence of substantial signal overlap, a combination of optimized isotope-labelling with isotope-filtered NOESY experiments was acquired to obtain unambiguous

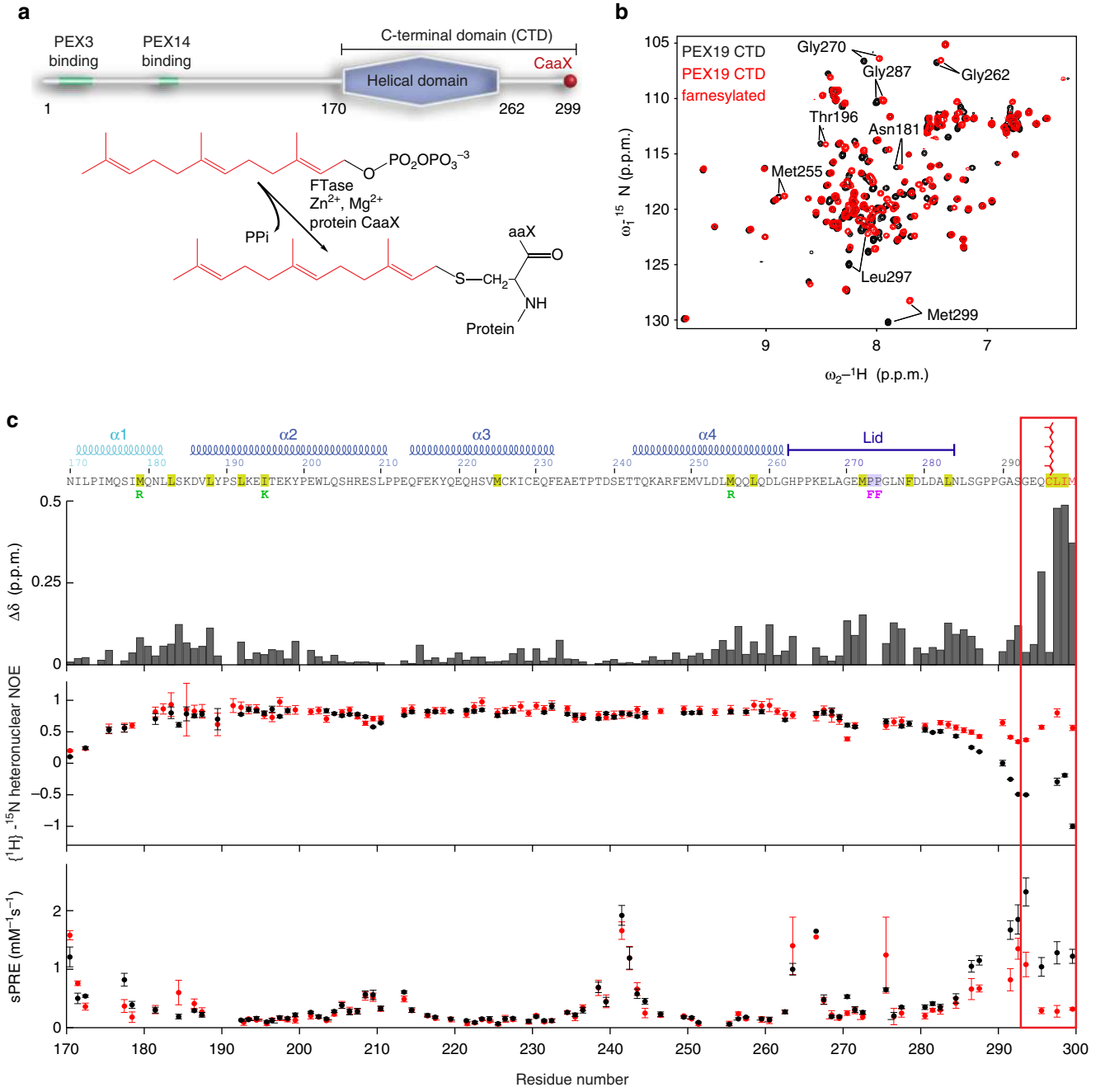


Figure 1 | NMR analysis of PEX19 farnesylation. (**a**) Schematic overview of human PEX19: the unfolded N-terminus of PEX19 interacts with the peroxisomal membrane proteins PEX3 and PEX14, the C-terminus harbours the α-helical cargo binding region and the farnesyl recognition sequence CaaX. Farnesyl transferase catalyses the transfer of the farnesyl moiety from farnesyl pyrophosphate to the cysteine of the CaaX box. (**b**) Overlay of $^{1}H,^{15}N$ HSQC spectra of PEX19 CTD with (red) and without (black) farnesylation. Amide resonances that are strongly affected by farnesylation are annotated. (**c**) Chemical shift perturbation (Δδ) of amide protons induced by farnesylation (top), $\{^{1}H\}$-^{15}N heteronuclear NOE (middle) and solvent paramagnetic relaxation enhancement (sPRE) rates for amide protons (bottom) are shown for PEX19 CTD with (red) and without (black) farnesylation. Error bars for heteronuclear NOE data based on s.d. of noise in the individual spectra were calculated using error propagation. Error bars for sPRE data represent fitting error of proton R_1 relaxation rates recorded from different concentrations of Gd(DTPA-BMA). The large changes in the C-terminal region that harbours the CaaX box are highlighted by a red box. The amino acid sequence and secondary structure elements of the PEX19 CTD are indicated on top, with aliphatic residues involved in farnesyl binding highlighted in yellow. Point mutations of residues that contact the farnesyl group or involved in PMP recognition are indicated as green and purple letters, respectively.

assignments of chemical shifts and NOEs (Supplementary Fig. 2a). The solution structure of the farnesylated PEX19 CTD was determined based on numerous NOE-derived distance restraints including 203 protein-farnesyl restraints and 136 residual dipolar couplings (Fig. 2; Supplementary Fig. 2b; Table 1) and was further validated by comparison of experimental and back-calculated solvent PREs for amide protons (Supplementary Fig. 2c). The structure of the farnesylated PEX19 CTD contains the N-terminal helix α1 followed by the three-helical bundle domain (helices α2, α3 and α4) and a linker that

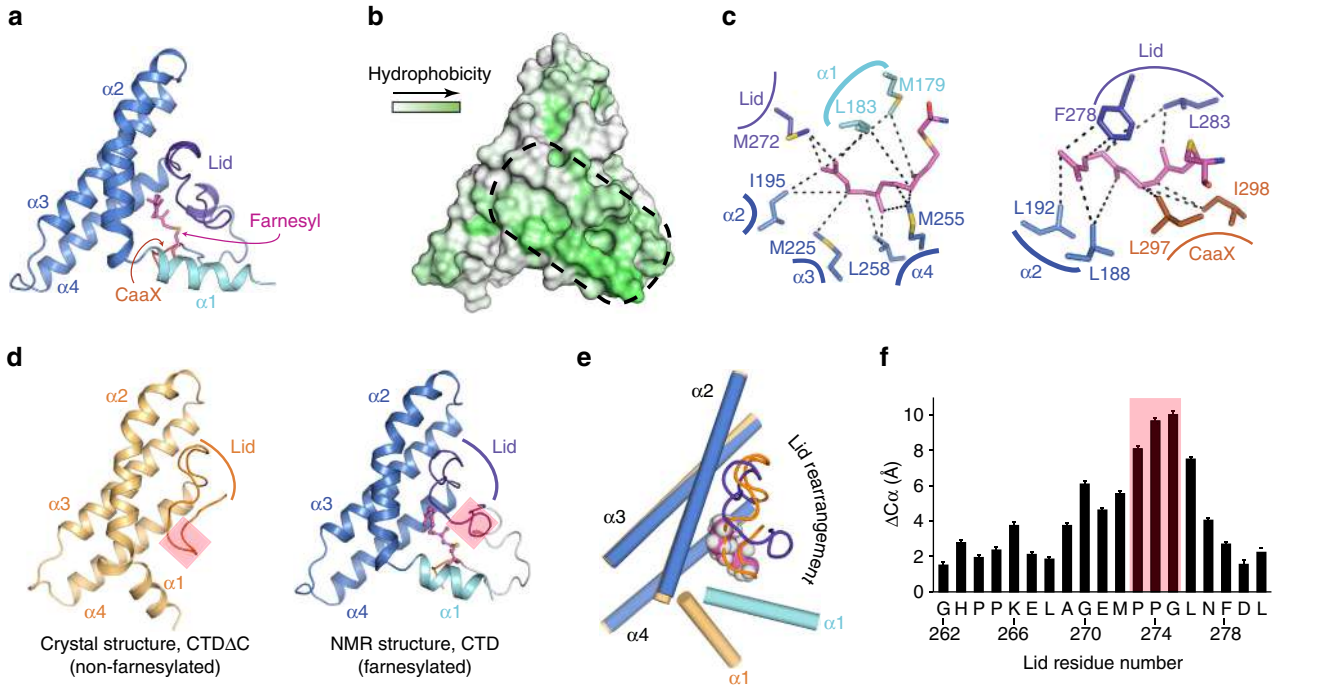


Figure 2 | Structure of farnesylated PEX19 CTD. (**a**) Cartoon representation of farnesylated PEX19 CTD. The farnesyl binding site is formed by residues in helix α1 (cyan), helices α2-α4 (blue), the lid region (residues 262–283; violet) and the CaaX box (orange). The farnesyl is drawn as sticks in magenta. (**b**) Surface representation of the farnesylated PEX19 CTD coloured by hydrophobicity. The hydrophobic surface region comprising helix α1 and the lid is indicated by a black dotted line. (**c**) Two different views of the farnesyl recognition site. Dashed lines indicate intermolecular NOE correlations, collected on PEX19 farnesylated samples using various labelling schemes on the protein side. The relative position of helices in the PEX19 CTD fold surrounding the farnesyl is indicated. (**d**) Comparison of the crystal structure of the PEX19 CTD lacking the C-terminal 16 residues (CTDΔC, left)[21], with the solution structure of the farnesylated PEX19 CTD (right). Structural changes of the lid region are highlighted by a red box. (**e**) The structure of the core helical bundle comprising helices α2-α4 is highly similar while the orientation of helix α1 (cyan) differs. The lid (residues 262–283) covers the farnesyl group in farnesylated PEX19 CTD (violet), while in the crystal structure of non-farnesylated PEX19 CTDΔC these residues occupy the farnesyl binding cavity (orange). Note, that residues 283–299 in the NMR structure and loop regions connecting the helices have been omitted for clarity. (**f**) Large conformational changes of the lid induced by PEX19 farnesylation are indicated by change in the Cα atom positions of lid residues with the helical core domain superimposed as in **e**. Error bars indicate the s.d. for the 20 models of the NMR ensemble. The largest changes are highlighted by a red box as in **d**.

connects α4 and the CaaX box (Fig. 2a). All four helices form a hydrophobic cavity that contacts the farnesyl group with aliphatic side chains: Met179, Leu183 in helix α1; Leu188, Leu192, Ile195 in helix α2; Met225 in helix α3; Met255, Leu258 in helix α4 (Fig. 2c). Residues within the linker region between α4 and the CaaX box (Met272, Phe278 and Leu283) mediate additional hydrophobic interactions to the farnesyl moiety and cover the binding cavity like a lid (Fig. 2a,c). The two aliphatic residues of the CaaX box (Leu297 and Ile298) contact the first isoprene unit and complete the binding pocket thus largely shielding the farnesyl group from the solvent. The farnesyl group is occluded in this internal cavity with a buried surface area of 457 Å^2 as determined by PDBePISA (ref. 22). The small size of the farnesyl binding cavity in the PEX19 CTD may thus stabilize a distinct bent conformation of the isoprenoid (Fig. 2; Supplementary Fig. 3).

The fold of the core three-helical bundle of farnesylated PEX19 CTD (helices α2, α3 and α4) is similar to the recently reported crystal structure of non-farnesylated PEX19 CTDΔC (ref. 21) (residues 161–283) (Fig. 2d) with a backbone coordinate r.m.s.d. of 0.84 ± 0.06 Å (residues 187–261). However, a number of important structural elements rearrange in a unique manner to form a continuous hydrophobic cavity and bury the farnesyl group, thus shielding it from the solvent. In the non-farnesylated protein the lid segment associates with other parts of the structure to shield hydrophobic patches and essentially occupies the position of the lipid (Fig. 2d,e). When compared to the non-farnesylated crystal structure of PEX19 CTDΔC, the most prominent differences involve a specific orientation of helix α1 to accommodate the bulky farnesyl group (Fig. 2e), and the displacement of the lid (Fig. 2f). Both elements form hydrophobic interactions with the farnesyl group and adopt a unique rigid arrangement relative to the three-helical bundle with the tip of the lid packing against the region between helices α1 and α2 (Fig. 2d). In the crystal structure of the non-farnesylated protein helix α1 is involved in a non-physiological tetramerization of the PEX19 CTDΔC (ref. 21), which defines its orientation. In contrast, the non-farnesylated PEX19 CTD (residues 161–299) is monomeric in solution (Supplementary Fig. 1c). Moreover, NMR relaxation data (Supplementary Fig. 1d) and the scarcity of intramolecular NOEs indicate that helix α1 is partially flexible and does not adopt a specific arrangement relative to the core helical fold. Thus, the structure of the farnesylated protein and our NMR data for the non-farnesylated protein demonstrate that farnesylation stabilizes a specific orientation of helix α1 and reshapes the lid region. Most importantly, the recognition of the farnesyl group in the internal cavity induces allosteric changes at the protein surface to form a hydrophobic groove (Fig. 2b) that serves as the binding site to PMP cargo peptides (see below).

Mutational analysis of farnesyl recognition. On the basis of the structure of farnesylated PEX19 CTD, we mutated residues that form the farnesyl binding pocket (M179R in helix α1, I195K in

Table 1 | NMR and refinement statistics for farnesylated PEX19.

NMR restraints	
Distance restraints	
Total no. of NOEs	3,728
Intra-residue	823
Inter-residue	2905
Sequential ($\|i-j\|=1$)	1,100
Medium-range ($\|i-j\|<4$)	1,002
Long-range ($\|i-j\|>5$)	600
Protein-farnesyl	203
Total no. of dihedral angle restraints	187
ϕ	91
ψ	96
Total no. of RDCs	136
Q^{RDC} (%)	28
Structure statistics	
Violations (mean ± s.d.)	
Distance restraints (Å)	0.018 ± 0.00
Dihedral angle restraints (°)	0.913 ± 0.06
Max. dihedral angle violation (°)	4.75
Max. distance restraint violation (Å)	0.46
Deviations from idealized geometry	
Bond lengths (Å)	0.003 ± 0.00
Bond angles (°)	0.511 ± 0.07
Impropers (°)	1.175 ± 0.04
Ramachandran plot statistics (%)	
Residues in most favored regions	91.5
Residues in additionally allowed regions	8.2
Residues in generously allowed regions	0.2
Residues in disallowed regions	0.1
Average pairwise r.m.s.d. (Å)*	
Backbone	0.45 ± 0.09
Heavy atoms	0.85 ± 0.11

*Pairwise coordinate r.m.s.d. was calculated for 20 refined structures for residues 175–285 and 295–299.

helix α2 and M255R in helix α4, Fig. 2c; Supplementary Table 1) to impair hydrophobic contacts and assess the functional importance of specific protein–lipid interactions. ^{1}H, ^{15}N NMR correlation spectra (Supplementary Fig. 4a) of the corresponding PEX19 CTD variants show that all mutant proteins studied maintain their structural integrity. Although these protein mutants are completely farnesylated *in vitro*, the introduction of the farnesyl group leads to differential chemical shift changes compared to the wild-type protein. Chemical shift differences of M179R and I195K versus the wild-type PEX19 CTD in the non-farnesylated and farnesylated state mainly involve residues adjacent to the mutation site, consistent with local effects introduced by replacing the corresponding residues (Supplementary Fig. 4b). In contrast, pronounced and wide-spread effects are seen for the M255R mutant, suggesting that the recognition and conformation of the farnesyl group in the PEX19 M255R variant is altered. To probe whether the M255R mutation leads to partial exclusion of the farnesyl moiety from the binding cavity, we compared the changes in $\{^{1}H\}$-^{15}N heteronuclear NOE values for the wild type and M255R PEX19 CTD upon farnesylation. Notably, while similar values are observed for the helical core domain, the farnesylation-induced increase of heteronuclear NOE values for the C-terminal residues seen for wild type PEX19 is less pronounced in the M255R PEX19 variant (Supplementary Fig. 4c). The reduced NOE values indicate an increased flexibility of the C-terminal region. The fact that the tail does not become completely flexible as seen for the non-farnesylated protein implies that the farnesyl group is not completely excluded from the cavity in the M255R PEX19 CTD variant. In any case, the significantly increased flexibility of the C-terminal region is consistent with a weakened protein-farnesyl interaction and suggests a partial and/or transient exposure of the hydrophobic farnesyl group in the M255R variant.

To provide further evidence for the (partial) solvent exposure of the farnesyl group in the mutant proteins, we employed hydrophobic interaction chromatography. Differences in surface hydrophobicity were determined for farnesylated and non-farnesylated wild type and mutant PEX19 CTD proteins (Fig. 3a; Supplementary Fig. 4d). As expected, the introduction of the farnesyl group generally increases the hydrophobicity of PEX19 CTD for both wild type and mutants. However, the change in hydrophobicity linked to farnesylation is significantly larger for all PEX19 CTD variants and increases from M179R to I195K to M255R. This is consistent with an increased exposure of the farnesyl group in these mutants, presumably due to steric clashes and/or charges introduced in the binding cavity. The hydrophobic interaction chromatography analysis shows the most pronounced effects for the M255R substitution consistent with a significant exposure of the farnesyl moiety in the M255R variant (Fig. 3a).

We next examined the role of farnesylation for the functional activity of PEX19. We transfected PEX19-deficient cells with a bicistronic vector expressing PEX19 variants and eGFP-SKL as peroxisomal targeting signal 1 (PTS1) marker. The PEX19-deficient fibroblasts are characterized by the absence of import-competent peroxisomes, which is indicated by the mislocalization of the peroxisomal matrix marker to the cytosol and mislocalization of the peroxisomal membrane marker PEX14 to mitochondria[23]. Complementation and thus reappearance of import-competent peroxisomes is indicated by an overlapping punctate pattern for eGFP-SKL and PEX14. Full-length wild type and PEX19 variants with mutations that affect the farnesyl interactions studied above were compared in a complementation assay in human *ΔPEX19* fibroblast cells. Whereas, wild type PEX19 restores peroxisomal protein transport to 60%, functional complementation was reduced to 46% in PEX19 C296S and further decreased to 23% for PEX19 I195K. The M179R and M255R mutations are not capable of rescuing the *ΔPEX19* phenotype (Fig. 3b,c). As shown by immunoblot analysis, wild type and mutant proteins are expressed in similar amounts and are farnesylated *in vivo*—with the exception of a C296S variant, which lacks the farnesylation site (Fig. 3d). Statistical analysis of the complementation efficiency reveals that peroxisomal protein import is partially impaired in the absence of farnesylation (C296S). Mutants that affect the recognition of the farnesyl group and the PMP cargo binding (I195K, M179R) strongly impair peroxisome formation. Notably, no rescue of peroxisome biogenesis is observed in a PEX19 mutant (M255R) that interferes with intramolecular farnesyl recognition, even though this residue is remote from the PMP binding surface. We also considered the possibility that the complementation defects of the PEX19 CTD mutants were caused by altered localization due to the exposure of the farnesyl moiety resulting in increased hydrophobicity. To this end, immunofluorescence microscopy revealed that import defects do not correlate with stronger association of PEX19 with peroxisomes or mislocalization to any other membrane-bound subcellular compartment (Supplementary Fig. 5).

Recognition of PMPs by farnesylated PEX19 CTD. Residues in helices α1 and α2 (residues 170–195) are highly conserved and have been previously implicated in binding of PMP cargo proteins[21]. Our structural analysis shows that farnesylation leads to a conformational rearrangement of this region and the lid in

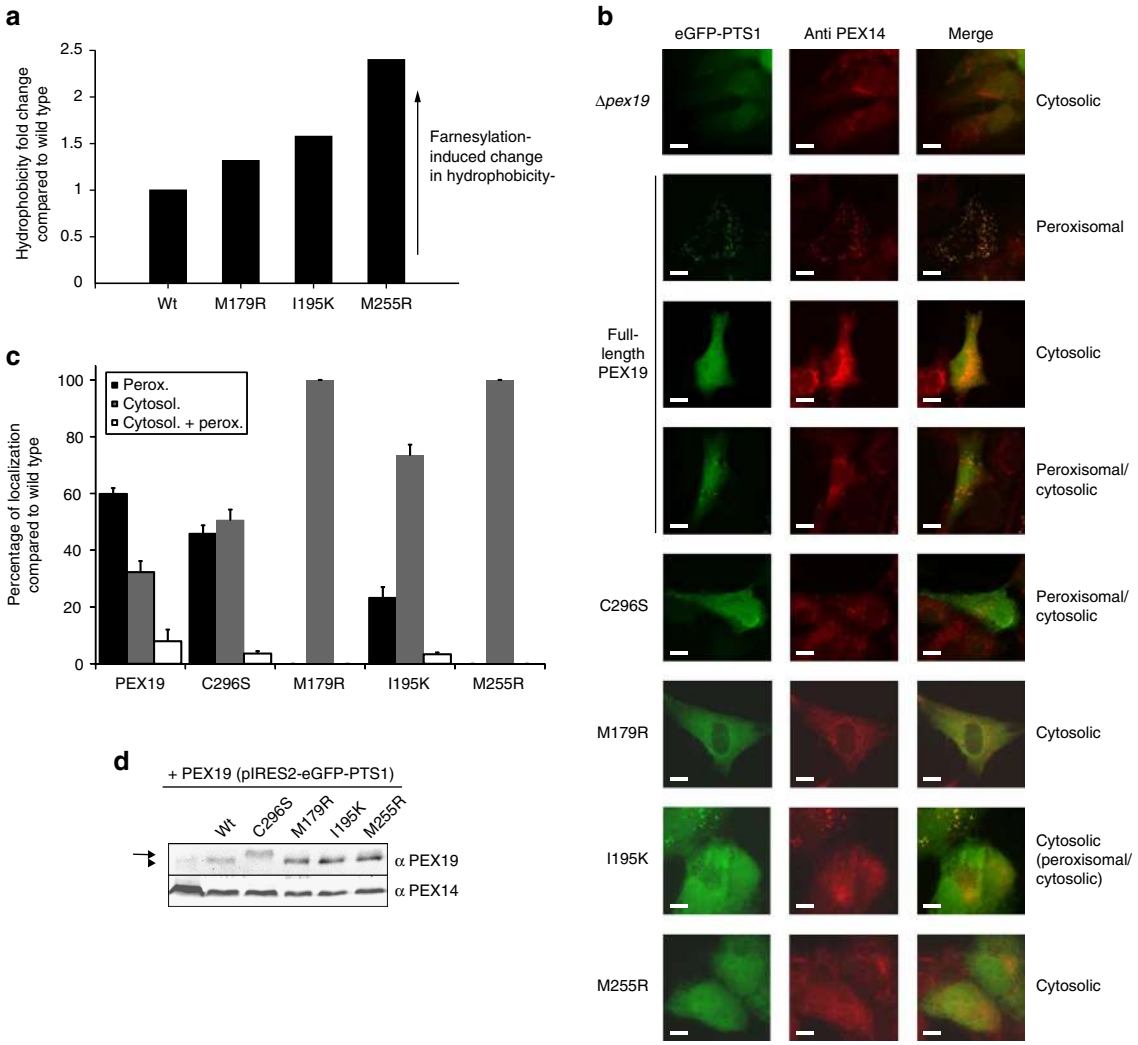


Figure 3 | Functional and biochemical analysis of PEX19 variants that affect farnesyl recognition. (**a**) Hydrophobic interaction chromatography analysis. Proteins were eluted from a Butyl Sepharose FF column using a linear gradient of decreasing $(NH_4)_2SO_4$ concentrations. For PEX19 CTD wild type and every variant, the farnesylated protein exhibits an increased elution volume compared to the non-farnesylated form (Supplementary Fig. 4d). The plot shows relative differences in hydrophobicity of the farnesylated and non-farnesylated state for each protein, compared to the wild type protein (which was set to 1). (**b**) Functional complementation of PEX19-deficient fibroblasts by PEX19 variants. PEX19 harbouring single-amino acid substitutions as indicated were introduced into PEX19-deficient fibroblasts by transfection with bicistronic expression vectors coding for full-length PEX19 variants and eGFP-PTS1. Localization of the model peroxisomal protein transport substrate eGFP-PTS1 was monitored by fluorescence microscopy (left column). Endogenous PEX14 as a peroxisomal membrane protein was detected using immunofluorescence microscopy (middle column). A punctate staining pattern and co-localization of eGFP-PTS1 and PEX14 indicate functional peroxisomal protein transport (merge, right column). For better comparison, PEX19-deficient cells expressing only eGFP-PTS1 from the vector are shown in the upper panels. Scale bar: 10 μm. The designation indicated on the right refers to the GFP-localization. (**c**) Quantitative analysis of Δ*PEX19*-phenotype complementation by the individual PEX19 variants. Values are obtained from analysing at least 100 cells in three independent transfection experiments. Data shown represent mean ± s.d. (**d**) Immunoblot analysis of the farnesylation mutants shows that all variants are expressed at wild type level and, except for C296S (arrow), are farnesylated *in vivo* (arrowhead).

PEX19 CTD. To map the binding surface with PMPs, we monitored NMR chemical shift perturbations upon addition of peptides that comprise previously characterized PEX19 binding regions[24] to ^{15}N-labelled farnesylated or non-farnesylated PEX19 CTD (Fig. 4). Probably, due to the hydrophobicity of the PMP-derived peptides their addition to PEX19 often resulted in precipitation. However, for a peptide derived from the ATP-binding cassette (ABC) transporter ALDP we could obtain soluble complexes at various PEX19-peptide ratios and analyse NMR chemical shift perturbations. Upon addition of the ALDP peptide to non-farnesylated PEX19 CTD, no significant spectral changes are observed at molar ratio up to 10:1 (Fig. 4a). In contrast, addition of the same peptide to farnesylated PEX19 CTD induces distinct chemical shift perturbations (Fig. 4b,c). The amide and methyl signals affected by the titration cluster to the hydrophobic groove that is formed on the surface of the farnesylated PEX19 structure (Figs 2b; 4d,e). The affinity for the ALDP peptide was measured by fluorescence polarization (Supplementary Fig. 6). For the wild type protein, farnesylation increased the affinity seven-fold, in agreement with findings reported for yeast PEX19 (ref. 16) and consistent with the NMR titration data. In contrast, all mutants (M179R, I195K, M255R) showed impaired binding with or without farnesylation (Supplementary Fig. 6).

We also compared the binding of other PMP-derived peptides to farnesylated and non-farnesylated PEX19 CTD using NMR

and microscale thermophoresis experiments. To this end, binding of a peptide derived from the PMP PEX13 induces even stronger NMR chemical shift perturbation changes to helix α1, and significant line broadening in the lid region, highlighting the importance of the lid in PMP binding (Supplementary Fig. 7a,b). For a PEX11-derived peptide we could quantitatively compare changes in binding affinity induced by farnesylation. These data show, that the interaction with non-farnesylated PEX19 CTD is reduced to a $K_D = 29.8 \pm 1.6\,\mu M$ compared to $K_D = 6.1 \pm 0.4\,\mu M$ for the farnesylated protein (Supplementary Fig. 7c).

The molecular recognition of the ALDP PMP peptide and farnesylated PEX19 CTD is documented by chemical shift perturbations and a large number of intermolecular NOEs observed mainly between methyl groups supporting further the methyl chemical shift perturbations (Fig. 4; Supplementary Fig. 8a, see methods for details). Few intermolecular NOEs could be assigned for the PEX19–PMP interaction involving methylene or methyl groups of PEX19 and the aromatic phenylalanine sidechain of the ALDP peptide. These intermolecular NOEs, chemical shift perturbations, and previously published mutational analysis[21] were used as ambiguous interaction restraints to calculate a structural model of the farnesylated PEX19 CTD–PMP peptide complex using HADDOCK (ref. 25). As the complex is not sufficiently stable to obtain assignments of the PEX19-bound form, the helical conformation of the docked peptide was inferred from secondary chemical shifts of the free peptide, which indicate a preformed helical conformation (Supplementary Fig. 8b). In fact most PMPs exhibit hydrophobic or aromatic residues spaced by three residues that are exposed on the same face of an α-helical conformation (Fig. 5a).

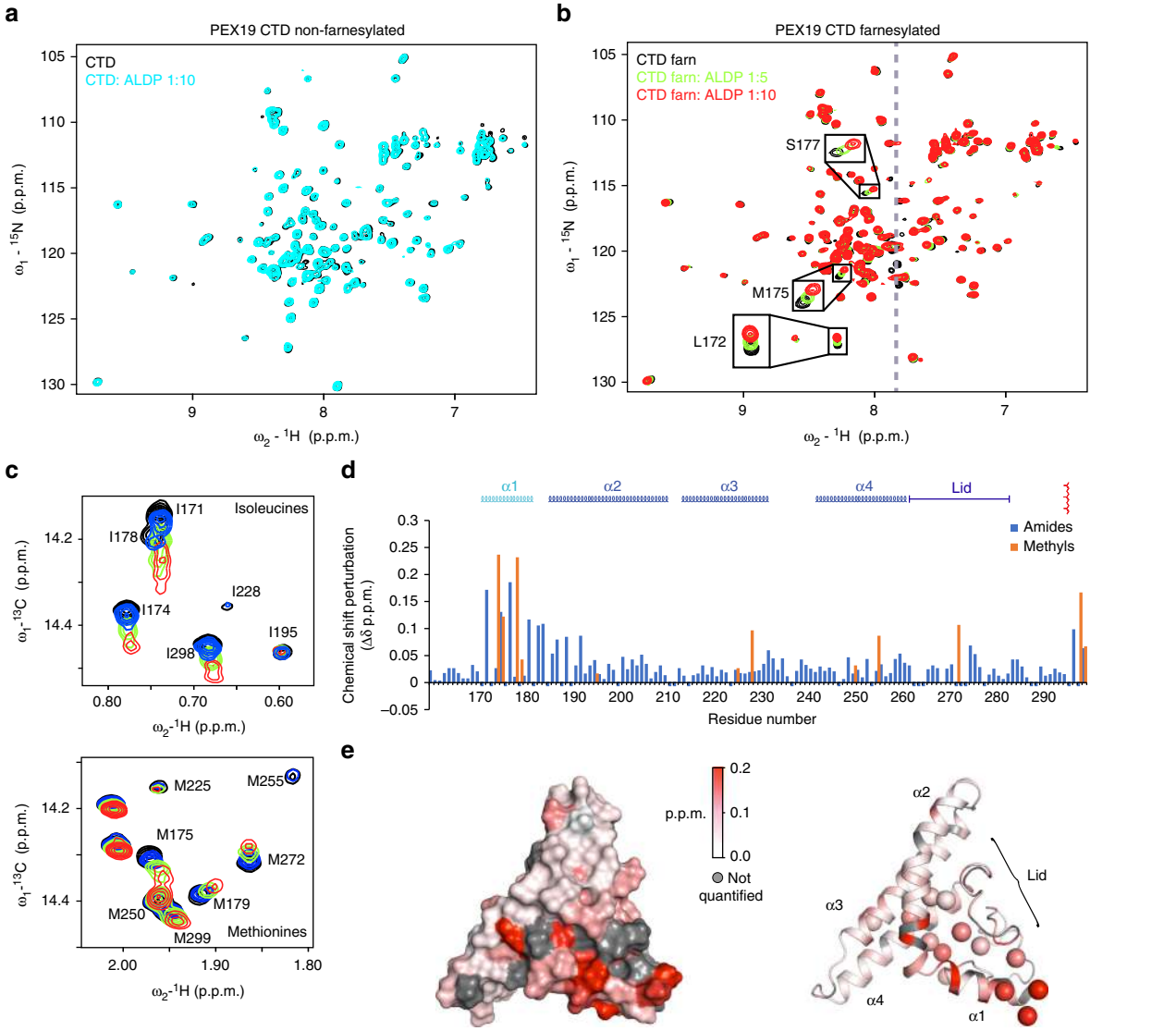
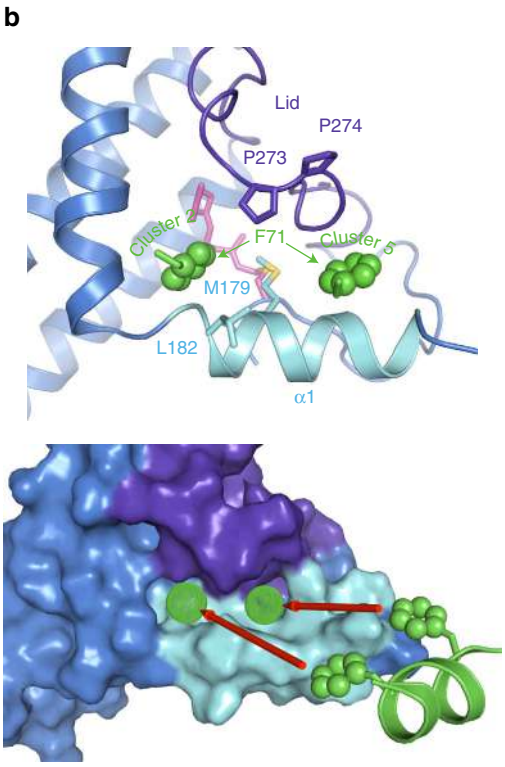


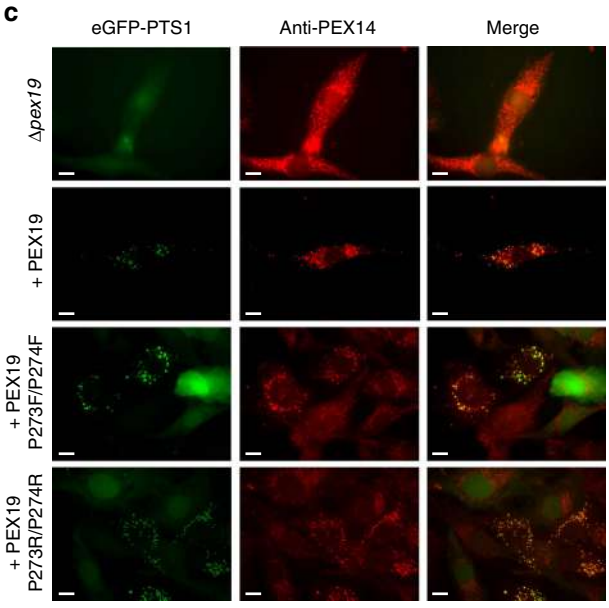
Figure 4 | NMR analysis of PMP binding to non-farnesylated and farnesylated PEX19 CTD. (**a**) Comparison of $^{1}H,^{15}N$ HSQC spectra of 100 μM non-farnesylated CTD free (black) and in the presence of 10-fold excess of the ALDP-derived peptide (cyan). (**b**) Comparison of $^{1}H,^{15}N$ HSQC spectra of 100 μM farnesylated PEX19 CTD free (black) and in the presence of 5 (green) and 10 (red) molar ratio of the ALDP peptide. The grey dashed line indicates the position of additional baseline correction applied to remove noise from DMF signals in the peptide stock solution. (**c**) Chemical shift perturbation of Met/Ile methyl groups seen in $^{1}H,^{13}C$ HSQC experiments upon increasing concentration of ALDP peptide. Black: free PEX19 CTD 100 μM, blue, green and red correspond to 200, 500 and 1,000 μM ALDP peptide concentration, respectively. (**d**) Quantification of chemical shift perturbation (CSP) of amide and methyl groups with respect to the PEX19 amino acid sequence. (**e**) Left: Surface representation of the PEX19 CTD coloured in red according to CSP. Right: cartoon representation with the backbone coloured based on amide CSP in red and spheres depicting methyl CSPs upon addition of the ALDP peptide (right). Grey colour refers to residues, which could not be analysed.

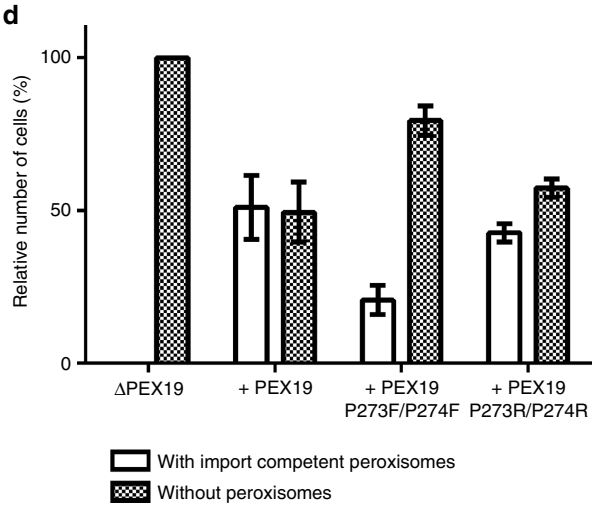
a

PMP	Position	Sequence
		....5...10...15...20...25.
ALDP*	61–76	AAAKAGMNRV**F**LQRLL**W**LLRLL**F**PRV
ALDPCext*	68–82	AAAKAGMNRV**F**LQRLL**W**LLRLL**F**PRV
PEX11a	173–198	**WF**SVAEEETE**W**LQS**F**LLLL**F**RSLKQH
PEX11b*	172–197	LALKLRLQVLLLARVLRGHPPLLLDV
PEX13*	174–189	ANH**F**SRLKIH**F**TKV**F**SA**F**ALVRTIR**Y**
PMP69	10–35	AGARPRLDLQ**F**LQR**F**LQILKVL**F**PS**W**
PEX12	58–72	PTH**Y**G**F**L**W**R**WF**DEI**F**TLLDLLLQQH**Y**

*Used in this study
Note: sequence in bold represent a known PEX19 binding peptide







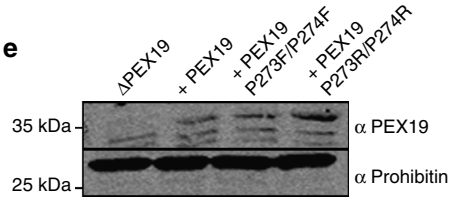


Figure 5 | Recognition of a PMP-derived peptide by the farnesylated PEX19 CTD. (**a**) Amino acid sequences of peroxisomal membrane proteins. Aromatic residues are highlighted in red; regions known to be involved in PEX19 binding are shown in bold. (**b**) Docking models showing two binding pockets for aromatic side chains of the PMP-derived peptide (green stick and ball representation). Top: The PEX19 lid and helix α1 are coloured dark blue and cyan, respectively. Side chains of M179, Leu182, Pro272 and Pro273 are shown as sticks. Bottom: Surface representation of the PEX19 CTD indicating the hydrophobic cavities that accommodate the aromatic side chain of Phe71 in the ALDP peptide. (**c**) *In vivo* effects of mutations of the PMP binding site of PEX19. Immunofluorescence microscopy images of PEX19-deficient human fibroblasts transfected with bicistronic vectors encoding for eGFP-PTS1 and different PEX19 variants. The same plasmid lacking PEX19 was used as a negative control (ΔPEX19) showing diffuse staining due to mitochondrial mislocalization for both peroxisomal marker proteins, eGFP-PTS1 (matrix protein) and PEX14 (PMP). The same plasmid lacking PEX19 was used as a negative control (ΔPEX19) showing cytosolic and mitochondrial mislocalization for the peroxisomal marker proteins eGFP-PTS1 (matrix protein) and PEX14 (PMP), respectively. A congruent punctate pattern of eGFP-SKL and PEX14 indicates formation of import-competent peroxisomes 72 h after transfection. Scale bar: 10 μm. (**d**) Statistical analysis of number of transfected cells containing import-competent peroxisomes. Values were obtained from 100 cells each of three independent transfected experiments. Data shown represent mean ± s.d. (**e**) Immunoblot analysis of mutants shows that all variants are present in an amount similar to the wild type protein.

The docking calculations yielded two clusters of structures, which differ with respect to the orientation of the ALDP peptide and the protein (Fig. 5b; Supplementary Table 2). This is explained by the scarcity of intermolecular restraints, which provide only distance information for the phenylalanine ring, and thus cannot unambiguously define the orientation of the PMP helix. Nonetheless, in both clusters an aromatic side chain occupies either of two hydrophobic pockets that are formed by the rearrangement of helix α1 and lid segment upon farnesylation (Figs 2e and 5b). This is consistent with previous mutational analysis[21] where disruption of helix α1 abolishes PMP binding by PEX19. However, our docking model also indicates that residues in the lid, that is, two proline residues (Pro273, Pro274), are involved in PMP recognition (Fig. 5b). Notably, the size and hydrophobic character of the PMP binding surface suffice to accommodate two hydrophobic residues of a PMP (Fig. 5b).

The physiological contribution of residues in the PMP binding site was probed by mutational analyses *in vivo*. Previous studies have already implicated residues in helix α1 in PMP binding *in vitro* and for peroxisome function in cells[21]. However, the role of residues in the lid has not been explored. Therefore, we studied the contributions of Pro273 and Pro274, two residues in the PEX19 lid that are implicated in ALDP/PMP binding (Fig. 5b; Supplementary Table 1) by substitution with arginine or phenylalanine. The phenylalanine substitutions were designed with the hypothesis that they would occupy the binding pockets for aromatic residues in the PMP cargo proteins and thus interfere with binding in *cis*. Full-length wild type and variant PEX19 proteins were expressed in PEX19-deficient human fibroblasts, which are characterized by the lack of peroxisomes. Cells that are not complemented by functional PEX19 show a cytosolic mislocalization of the peroxisomal marker protein eGFP-PTS1 and a diffuse background immunolabelling of the PMP PEX14, which is distinct from the peroxisomal pattern and known to reflect mislocalization to mitochondria (Fig. 5c, ΔPEX19). After 72 h transfection, 50% of cells expressing wild type PEX19 display formation of newly import-competent peroxisomes as indicated by the congruent punctate pattern of both marker proteins (Fig. 5b, +PEX19). Cells, which express the PEX19 P273F/P274F mutant, displayed a significantly lower complementation rate of only 20% (Fig. 5d), independent of the PEX19 protein level (Fig. 5e). The impaired function of the PEX19 mutant demonstrates that the lid region is not only important for binding of the half-transporter ALDP but also for other PMPs, in particular for those which are essential for biogenesis of peroxisomes.

Discussion

Our structural analysis shows that the farnesyl group in the PEX19 CTD is buried within a hydrophobic cavity. Note, that this cavity has been found to be empty in the previously reported crystal structure[21]. Consistent with this, substantial conformational differences are seen in the farnesylated PEX19 CTD when compared to the crystal structure of the non-farnesylated PEX19 CTDΔC (ref. 21). An unexpected finding of or structural analysis is that PEX19 farnesylation stabilizes a distinct conformation of helix α1 and the C-terminal region of PEX19 and thereby arranges the PMP binding site, suggesting an allosteric modulation of the PMP interaction. Farnesylation affects PMP binding by preorganization of the PMP binding surface with an appropriate placement of the helix α1 and the lid and thereby provides a hydrophobic binding surface for PMP binding. The functional significance of this allosteric control of PMP binding is confirmed by our mutational analysis *in vitro* and *in vivo,* which shows that PEX19 mutations that affect farnesyl recognition remote from the PMP binding site strongly interfere with peroxisome function.

To our knowledge, the structure of the farnesylated human PEX19 CTD is the first example of a protein, which buries the farnesyl group by an internal cavity formed by the same polypeptide chain *in cis*. A recent structural analysis of a farnesylated Ras protein found that the farnesyl group is recognized by the guanine nucleotide dissociation inhibitor (GDI) PDEδ *in trans* and is released in an allosteric manner by the G proteins Arl2/Arl3 (refs 26,27). Such an allosteric control is reminiscent of the modulation of PMP binding by PEX19 by farnesylation *in cis* as observed in the present study.

Other structurally well-characterized examples of isoprenoid binding are farnesyltransferases (Supplementary Fig. 3). In these enzymes, conserved aromatic amino acids interact with the isoprene units of the substrate, thereby recognizing the farnesyl in an extended conformation[28]. Farnesyl recognition by PEX19 is distinct from farnesyl transferase complexes, as mainly aliphatic residues contact the farnesyl, which adopts a bent conformation. The bent conformation of the farnesyl group is presumably linked to the small size of the PEX19 fold (Supplementary Fig. 3).

Recent studies have shown that non-farnesylated PEX19 recognizes PMP cargo proteins less efficiently. Accordingly, *S. cerevisiae* cells lacking the farnesylation motif are deficient in growth on oleic acid as sole carbon source, a metabolic condition that depends on functional peroxisomes[16]. Here, we show that farnesylation of the PEX19 CTD leads to an enhanced PMP binding. We note that additional effects of PEX19 farnesylation could contribute to the functional activity of the PEX19 protein. (1) The presence of a farnesyl group may influence the PMP membrane insertion, that is, farnesyl exposure could support PMP cargo release by interactions with the membrane, PEX3 or other factors. (2) It is also important to note that farnesylation may be only a first step of posttranslational modification of PEX19. It is conceivable that the farnesylated CaaX box is further processed by C-terminal proteolysis and carboxymethylation. This modification could be a way to regulate the activity of PEX19, as the carboxymethylation will further enhance the hydrophobicity of the C-terminal region and thus modulate interactions at the membrane. The presence of this PEX19 modification in cells has not yet been experimentally determined and its functional role should be investigated in future studies.

The structural analysis shows that a specific conformation of the conserved PMP binding site is assembled and controlled by farnesylation of the PEX19 CTD. Our NMR data and docking calculations demonstrate that the PMP binding maps to the region comprising helix α1 and the lid region[21], involving substantial structural rearrangements (Fig. 2e,f). The groove formed on the surface by the lid and helix α1 is sufficiently large to accommodate hydrophobic side chains that are spaced by about three amino acids, that is, corresponding to a single helical turn in the PMP ligand peptides. This is fully consistent with the hydrophobic nature of PMP peptides and an α-helical conformation that—as observed for the ALDP peptide—is partially preformed, already in the absence of PEX19. The observation of two cavities, which can accommodate aromatic and/or aliphatic side chains exposed from a helical peptide suggests optimal features of membrane protein peroxisomal targeting signals (mPTS) in peroxisomal membrane proteins. Indeed all PMPs exhibit hydrophobic mPTS motifs, which are predicted to adopt an α-helical conformation and comprise aromatic or aliphatic side chains in a spacing of 3–4 residues.

For the recognition of PMPs, helix α1 can be thought of as the thumb and the lid as the fingers of a hand, which, upon farnesylation, grasps the PMP (Fig. 6). Previously identified mutations on helix α1, which probably disrupt the helix, have

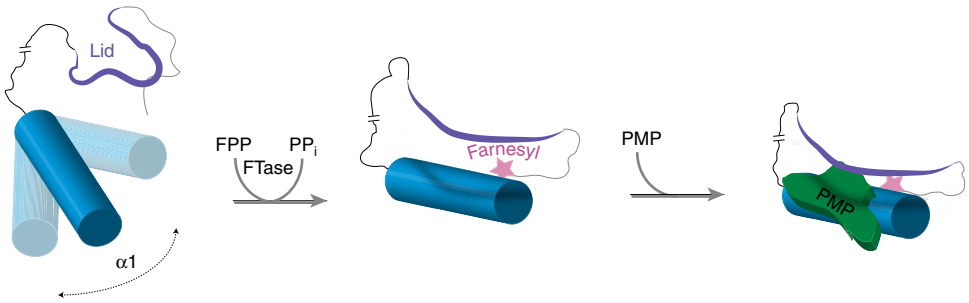


Figure 6 | Schematic model for the allosteric regulation of PMP binding by farnesylation. In case of non-farnesylated PEX19 CTD, the helix α1 is flexible in solution and only loosely attached to the helical core domain of the PEX19 CTD, while the lid adopts an open conformation. Farnesylation induces a specific, rigid arrangement of these two structural elements, thus forming a hydrophobic cavity, which then forms a high-affinity binding site for PMPs.

been shown to strongly affect the PMP binding[21]. Here, we provide direct structural evidence that both helix α1 and the lid, which is stabilized in a unique conformation by farnesylation of the PEX19 CTD, provide key interactions for PMP binding. In particular, replacement of two consecutive prolines in the lid by phenylalanine interferes with PMP binding and therefore affects peroxisomal biogenesis.

Our mutational analysis shows that mutation of M255, which is important for farnesyl recognition but remote from the PMP binding site, significantly impairs the PMP interaction (Fig. 3). This strongly argues that farnesylation stabilizes a specific conformation of the binding surface and thereby enhances the binding affinity for PMP cargo, as we have determined for different PMP-derived peptides in the present study. The structural rearrangements coupled to PEX19 farnesylation therefore suggest an allosteric control of PMP binding by farnesylation (Fig. 5).

To investigate the functional importance of the farnesyl recognition by PEX19, we monitored the functional activity of mutant proteins that impair farnesyl binding. The mutations introduced may influence PMP binding in two ways.

Mutations may affect direct contacts with the PMP cargo and thereby reduce functional activity. Schueller *et al* have previously shown that mutations of conserved hydrophobic residues in helix α1, such as Ile178 and Leu182, or truncation of this helix abolish PMP binding and the functional activity of PEX19 (ref. 21). Consistent with these observations, the M179R variant in helix α1 investigated in the present study cannot restore peroxisomal biogenesis in *ΔPEX19* fibroblasts. This mutation presumably destabilizes the orientation of helix α1 and consequently disturbs the formation of the hydrophobic pocket (see NMR chemical shift perturbations, Fig. 5b; Supplementary Fig. 4b). Here, we demonstrate that the binding site also involves the lid, as proline residues in the lid have an equally important contribution to the binding (Figs 5b and 6).

On the other hand, mutations may impair functional activity by affecting PMP binding indirectly via allosteric conformational changes induced at the PMP binding surface. The role of farnesylation is demonstrated by the PEX19 C296S variant, which is not farnesylated and cannot fully restore PEX19 function. The functional activity of the PEX19 I195K variant is also impaired although less strongly, presumably as Ile195 contacts the distal part of the farnesyl group and thus has smaller contributions to farnesyl recognition. In contrast, the M255R variant completely abolishes PEX19 function in fibroblasts, even though the mutant protein is expressed and farnesylated *in vivo*. This mutation induces extensive chemical shift perturbations in the PEX19 CTD compared to the wild type protein (Supplementary Fig. 4b), suggesting that the recognition of the farnesyl group in the internal cavity is strongly perturbed. Enhanced hydrophobicity associated with partial exposure of the farnesyl group is indicated by hydrophobic interaction chromatography (Fig. 3a; Supplementary Fig. 4d). These data demonstrate that proper recognition of the farnesyl group by burial in the internal cavity of PEX19 plays an important role for its functional activity *in vivo*.

In conclusion, our study reveals an unprecedented mode of farnesyl recognition *in cis* by the conserved α-helical fold of the PEX19 CTD. The insight into structural and functional effects of PEX19 farnesylation represents a significant advance towards mechanistic understanding of the role of PEX19 in peroxisomal protein transport. The findings observed here for PEX19 suggest that protein farnesylation may not only affect membrane localization of proteins but may have more general and novel roles in the regulation of protein function by allosteric control of ligand binding.

Methods

Expression and protein purification. *Escherichia coli* BL21 (DE3) (Novagen) was used for protein expression. His-tagged human PEX19 was expressed from pETM11 vector and purified as soluble protein using Ni^{2+} affinity chromatography (nickel-nitrilotriacetic acid)[21]. Cell pellets were resuspended and sonicated in 30 mM Tris–HCl, pH 8.0, 300 mM NaCl, 10 mM imidazole and 1 mM β-mercaptoethanol. Lysate was centrifuged at 15,000*g* for 20 min, and supernatant was loaded on Ni-NTA column. Subsequently the resin was washed with 30 mM Tris–HCl, pH 8.0, 700 mM NaCl, 25 mM imidazole, 1 mM β-mercaptoethanol and protein was eluted with 30 mM Tris–HCl, pH 8.0, 150 mM NaCl, 400 mM imidazole, 1 mM β-mercaptoethanol. Hexahistidine tag was cleaved overnight by TEV protease and removed by additional Ni^{2+} affinity chromatography. Flow through was loaded on size-exclusion chromatography column for final purification step. Uniformly isotope-labelled samples were obtained using M9 medium supplemented with ^{13}C-labelled glucose and/or $^{15}NH_4Cl$, respectively. For perdeuterated proteins, M9 medium in D_2O and [U- ^{2}H]-D-glucose were used for protein expression. *S. cerevisiae* farnesyl transferase (FTase) was expressed from a bicistronic pETM11 vector (kindly provided by S. Holton, EMBL Hamburg) and purified as described for PEX19 without removing the hexahistidine tag.

***In vitro* farnesylation.** *In vitro* farnesylation assays were carried out as described[29,30] in 50 mM Tris–HCl, pH 8.0, 20 mM KCl, 5 mM $MgCl_2$, 10 μM $ZnCl_2$ and 10 mM dithiothreitol. A reaction volume of 10 ml with 100 μM PEX19 was incubated with 130 μM farnesyl pyrophosphate (Sigma Aldrich) and 250 nM farnesyl transferase for 1 h at 37 °C. A subsequent Ni^{2+} affinity chromatography (nickel-nitrilotriacetic acid) removed the FTase. The farnesylated PEX19 was further purified by size exclusion chromatography on a HiLoad 16/60 Superdex 75 (GE Healthcare) and stored in 20 mM potassium phosphate, pH 6.5 and 50 mM NaCl. The completeness of the farnesylation was analysed by SDS gel electrophoresis and ^{1}H,^{15}N HSQC NMR spectra.

NMR Spectroscopy. NMR spectra were acquired at 298 K on a Bruker Avance III 600-MHz spectrometer with a TCI cryo-probe head, an Avance III 750-MHz spectrometer with a TXI probe head, or an Avance 900 instrument with a TXI cryo-probe head. Spectra were processed using NMRPipe[31] and analysed with Sparky[32]. Backbone assignment was done semi-automatically with MARS[33].

Backbone and side chain assignments were obtained from HNCA, HNCACB, CBCA(CO)NH, H(C)CH-TOCSY and (H)CCH-TOCSY spectra[34]. ^{13}C- and ^{15}N- edited NOESY spectra with a mixing time of 70 ms were recorded to derive distance restraints. For NOEs between PEX19 and the farnesyl group, isotope-edited and filtered 3D NOESY spectra (70 and 100 ms mixing times) were acquired using specific isotope-labelled samples. ^{15}N R_1, R_2 relaxation rates and $\{^1H\}$-^{15}N heteronuclear NOEs were recorded at an Avance III 750 MHz spectrometer[35]. Local correlation times were derived from the ^{15}N R_2/R_1 ratio[35]. For determination of solvent paramagnetic relaxation enhancement rates, gadolinium diethylene-triamine-penta-acetic-acid bismethylamide (Gd(DTPA-BMA)) was titrated to the protein samples[36–38]. Saturation recovery ^{1}H,^{15}N HSQC experiments with recovery times from 0.01 to 3 s were recorded at concentrations of 0, 1, 2, 3, 5, 7 and 10 mM Gd(DTPA-BMA). Peak intensities were fitted to an exponential recovery function according to equation $I(\tau) = I_0 - A\exp(-R_1\tau)$, where τ is the recovery delay, $I(\tau)$ is the peak intensity measured for the recovery delay τ, I_0 is the maximum peak intensity, R_1 is the longitudinal proton relaxation rate and A is the signal amplitude. To obtain the sPRE value, the proton R_1 rates were collected for all measured concentrations of Gd(DTPA-BMA) c_{para} and a linear regression for the equation R_1 (c_{para}) $= m_{sPRE}\, c_{para} + R_1^0$ was performed, where R_1 (c_{para}) is the proton R_1 measured at the concentration c_{para} of the paramagnetic compound, the slope m_{sPRE} corresponds to the sPRE and R_1^0 is the fitted proton R_1 in the absence of the paramagnetic compound[36]. For NMR titrations an ALDP peptide (sequence: AAKAGMNRVFLQRLL, residues 62–76 of human ALDP) was dissolved to a concentration of 5 mM in NMR buffer and added to ^{15}N-labelled PEX19 CTD with and without farnesylation in the same buffer up to a protein-peptide molar ratio of 1:10.

Structure calculation. NOE cross-peak assignments and structure calculations with torsion angle dynamics were carried out automatically with CYANA 3.0 followed by manual inspection[39]. The farnesylated cysteine was parametrized by using the Dundee PRODRG2 Server[40]. The CYANA library was prepared manually using the PDB coordinates derived from the PRODRG2 Server and all rotatable dihedral angles were defined. The dihedral angles for the isoprene double bonds were fixed to *trans* in accordance with our NMR analysis that defined the lipid as (2*E*,6*E*)-farnesol. For the last isoprenoid unit the methyls were stereospecifically assigned based on their carbon chemical shifts. Number 14 is the *trans* methyl as it has an upfield ^{13}C chemical shift that is in the same frequency range as found for the ^{13}C chemical shifts of the *trans* methyls 4 and 10 ($\sim$20 p.p.m.). Number 15 is the *cis* methyl based on its downfield ^{13}C chemical shift ($\sim$28 p.p.m.). For farnesyl atom numbering please see Supplementary Fig. 2a. No farnesyl-farnesyl restraints were imposed during the structure calculations. Protein-protein and protein-farnesyl NOE peak intensities were converted to upper distance limits using the internal calibration function of CYANA. Unambiguous distance restraints derived from CYANA, backbone dihedral angle restraints derived from TALOS$^+$ (ref. 41), and residual dipolar coupling restraints were used for water refinement using CNS (ref. 42). For water refinement the parameter and topology files were taken from the PRODRG2 Server and stereochemistry was the same as in CYANA. All structures were validated with iCing (http://nmr.cmbi.ru.nl/icing/). Molecular images were generated with PyMol (Schrödinger).

Intermolecular NOEs. Based on initial ^{13}C-edited, ^{13}C,^{15}N filtered NOESY experiments we were able to identify the methylene groups of the prolines and methyl groups in contact with the phenylalanine ring of ALDP peptide. The aromatic frequencies could be unambiguously assigned to ALDP peptide, because they were unique. Strong NOEs were also observed for ALDP methyl frequencies (Supplementary Fig. 8a) but could not be unambiguously assigned due to the large number of methyl-containing residues in the ALDP peptide. Nevertheless, these NOEs confirm that numerous contacts, involving the aromatic residues and methyl groups in the PMP ligand establish a specific complex with farnesylated PEX19. Assignments of NOEs to other ALDP residues were not possible due to overlap with residues in the farnesyl group, which was not isotope-labelled. To improve sensitivity and reduce spectral overlap PEX19 CTD was uniformly ^{2}H,^{13}C, ^{15}N-labelled, except for Leu and Val methyl groups. Intermolecular NOEs were measured for complexes of this PEX19 CTD bound to the unlabelled ALDP peptide using ^{13}C-edited NOESY experiments[34]. From these experiments proton chemical shifts at $\approx$0.9 p.p.m. were unambiguously assigned to Leu182 or Leu183 (Supplementary Fig. 8a).

Docking calculations. On the basis of NMR titrations and NOESY experiments both ambiguous and unambiguous restraints were obtained for the complex PEX19 CTD-ALDP. The expert interface of HADDOCK docking server[25] was used to perform docking calculations. The NMR structure of farnesylated PEX19 CTD, and a helical fragment of the ALDP peptide (residues 68–76: NRVFLQRLL) were used as first and second molecule respectively. PEX19 residues with CSP > 0.05 p.p.m. (Leu172, Met175, Ser177, Asn181, Leu182, Leu183, Ser184, Lys185, Y189, L192) and the Phe71 in the ALDP peptide were used as active residues and those within 6.5 Å were treated as passive residues. In iterations 0, 1 and during water refinement 2,000, 400 and 400 structures were calculated, respectively. The minimum cluster size was set to 10. Ambiguous distance restraints were derived from ω_1-filtered/ω_2-edited NOESY experiment and involve methyl protons of PEX19 Leu182 or Leu183 and methylene protons of Pro273 or Pro274 to ALDP aromatic protons of Phe71.

Microscale thermophoresis assays. An N-terminally fluorescein-labelled peptide LALKLRLQVLLLARV, corresponding to residues 186–200 of human PEX11B (Peptide Specialty Laboratories GmbH, Heidelberg) was dissolved to a concentration of 200 nM in 20 mM potassium phosphate, pH 6.5, 50 mM NaCl with 20% methanol and 0.1% Tween. Unlabelled PEX19 CTD with or without farnesylation was serially diluted in the same buffer to concentrations from 612 µM to 18 nM for wild type PEX19 CTD. Equal volumes of protein and peptide solutions were mixed thoroughly and centrifuged at 15,000*g* for 5 min. The supernatant was transferred to Microscale Thermophoresis Assays (MST) capillaries and measured with a NanoTemper Monolith NT 0.15T. Normalized fluorescence values from three separate measurements for PEX19 CTD wild type were used to determine the dissociation constants.

Fluorescence polarization. ALDP peptide (62-AAKAGMNRVFLQRLL-76) fused N-terminally with fluorescein isothiocyanate (FITC), was purchased from Peptide Specialty Laboratories GmbH. The lyophilized powder was dissolved in 1 ml water resulting in a 5 mM stock solution. Overnight dialysis against buffer containing 50 mM Tris pH 8.0, 50 mM NaCl, removed trifluoroacetate counterions. For the determination of each binding curve a 12 point titration was performed with constant concentration of 20 nM labelled peptide with increasing concentrations of PEX19 CTD variants. Starting concentration of PEX19 CTD was 1 mM with 12 serial dilutions reaching 0.5 µM. Reaction mixtures were then transferred into 96-well *Optiplate* and measured as triplicates in PerkinElmer EnVision plate reader.

Hydrophobic interaction chromatography. Purified PEX19 CTD and variants were adjusted to a buffer containing 650 mM $(NH_4)_2SO_4$ and 500 µg protein was subjected to equilibrated Butyl Sepharose FF column (20 mM potassium phosphate, pH 6.5, 50 mM NaCl and 650 mM $(NH_4)_2SO_4$). After washing, bound protein was eluted with linear decreasing $(NH_4)_2SO_4$ concentrations (650–0 mM $(NH_4)_2SO_4$, 20 ml). Eluted proteins were detected using 280 nm absorption. The binding strength is indicated by the elution volume and correlates with protein hydrophobicity.

Peroxisomal protein import assays in fibroblasts. A bicistronic expression plasmid coding for eGFP-PTS1 and non-tagged PEX14 (pMF1220)[43] was kindly provided by Marc Fransen (Leuven, Belgium). pIRES2-HsPEX19-eGFP-PTS1 was constructed by replacing the open reading frame of PEX14 (BglII/SalI) with a DNA-cassette encoding the full-length open reading frame of human PEX19 as derived from BamHI/SalI digestion of pAH05 (ref. 44). Single point mutations were introduced into the PEX19 sequence by Quickchange XL Site-directed Mutagenesis kit (Stratagene) using primer pairs (Supplementary Table 1). The bicistronic expression vectors coding for PEX19 variants and eGFP-PTS1 were transfected into the human cell-line ΔPEX19 T which was derived from PEX19-deficient Zellweger patient fibroblasts[21]. ΔPEX19 T cells, characterized by a defect in peroxisome biogenesis were routinely authenticated by functional complementation assays as described below and checked for mycoplasma contamination using Venor GeM Advance Mycoplasm Detection Kit (Minerva Biolabs). The efficiency of transfection according to the Amaxa Cell Line Nucleofector Kit R protocol (Lonza Cologne AG) was always around 90%. After 72 h of transfection, cells were subjected to fluorescence and immunofluorescence microscopy using polyclonal antiserum against human PEX14 (ref. 45) (1:400 diluted) or monoclonal antibodies against human PEX19 (BD Biosciences, MA5-17266 Thermo Scientific; 1:400 diluted). Statistical analysis was carried out from three independent transfections of each pIRES-PEX19-eGFP-PTS1 expression plasmid. Based on the appearance of eGFP-PTS1 fluorescence pattern, about 100 cells of each experiment were visually categorized into three classes, (i) full complementation of peroxisomal import as visualized by punctuate staining pattern, (ii) partial complementation as indicated by diffuse cytosolic staining with few dots, (iii) no complementation resulting in only cytosolic staining. All micrographs were recorded on a Zeiss Axioplan 2 microscope with a Zeiss Plan-Apochromat $\times$63/1.4 oil objective and an Axiocam MR digital camera, and were processed with AxioVision 4.6 software (Zeiss, Jena, Germany). The steady-state level of PEX19 expression was assessed by immunoblot analysis, using monoclonal mouse antibodies against HsPEX19 (BD Biosciences, MA5-17266; 1:1,000 diluted) (Supplementary Fig. 9).

Data availability. UniProt accession numbers used in this study are P40855 (PEX19), P22007/P29703 (farnesyl transferase), Q92968 (PEX13), O96011 (PEX11B) and P33897 (ALDP). The atomic coordinates and restraint files for the NMR ensemble of the farnesylated PEX19 CTD are deposited in the PDB with accession code 5LNF, chemical shifts are deposited at BMRB, entry 34030. All data are available from the authors.

References

1. Steinberg, S. J. *et al.* Peroxisome biogenesis disorders. *Biochim. Biophys. Acta* **1763,** 1733–1748 (2006).
2. Rucktaschel, R., Girzalsky, W. & Erdmann, R. Protein import machineries of peroxisomes. *Biochim. Biophys. Acta* **1808,** 892–900 (2011).
3. Emmanouilidis, L., Gopalswamy, M., Passon, D. M., Wilmanns, M. & Sattler, M. Structural biology of the import pathways of peroxisomal matrix proteins. *Biochim. Biophys. Acta* **1863,** 804–813 (2016).
4. Meinecke, M. *et al.* The peroxisomal importomer constitutes a large and highly dynamic pore. *Nat. Cell Biol.* **12,** 273–277 (2010).
5. Jones, J. M., Morrell, J. C. & Gould, S. J. Pex19 is a predominantly cytosolic chaperone and import receptor for class 1 peroxisomal membrane proteins. *J. Cell Biol.* **164,** 57–67 (2004).
6. Fang, Y., Morrell, J. C., Jones, J. M. & Gould, S. J. Pex3 functions as a pex19 docking factor in the import of class i peroxisomal membrane proteins. *J. Cell Biol.* **164,** 863–875 (2004).
7. Matsuzono, Y. & Fujiki, Y. *In vitro* transport of membrane proteins to peroxisomes by shuttling receptor pex19p. *J. Biol. Chem.* **281,** 36–42 (2006).
8. Chen, Y. *et al.* Hydrophobic handoff for direct delivery of peroxisome tail-anchored proteins. *Nat. Commun.* **5,** 5790 (2014).
9. Fujiki, Y., Matsuzono, Y., Matsuzaki, T. & Fransen, M. Import of peroxisomal membrane proteins: the interplay of pex3p- and pex19p-mediated interactions. *Biochim. Biophys. Acta* **1763,** 1639–1646 (2006).
10. Banerjee, S. K., Kessler, P. S., Saveria, T. & Parsons, M. Identification of trypanosomatid pex19: Functional characterization reveals impact on cell growth and glycosome size and number. *Mol. Biochem. Parasitol.* **142,** 47–55 (2005).
11. Manne, V. *et al.* Identification and preliminary characterization of protein-cysteine farnesyltransferase. *Proc. Natl Acad. Sci. USA* **87,** 7541–7545 (1990).
12. Magee, T. & Seabra, M. C. Fatty acylation and prenylation of proteins: what's hot in fat. *Curr. Opin. Cell Biol.* **17,** 190–196 (2005).
13. Hancock, J. F. Ras proteins: different signals from different locations. *Nat. Rev. Mol. Cell Biol.* **4,** 373–384 (2003).
14. Kammerer, S. *et al.* Genomic organization and molecular characterization of a gene encoding hspxf, a human peroxisomal farnesylated protein. *Genomics* **45,** 200–210 (1997).
15. Matsuzono, Y. *et al.* Human pex19: Cdna cloning by functional complementation, mutation analysis in a patient with zellweger syndrome, and potential role in peroxisomal membrane assembly. *Proc. Natl. Acad. Sci. USA* **96,** 2116–2121 (1999).
16. Rucktaschel, R. *et al.* Farnesylation of pex19p is required for its structural integrity and function in peroxisome biogenesis. *J. Biol. Chem.* **284,** 20885–20896 (2009).
17. Sato, Y. *et al.* Structural basis for docking of peroxisomal membrane protein carrier pex19p onto its receptor pex3p. *EMBO J.* **29,** 4083–4093 (2010).
18. Schmidt, F. *et al.* Insights into peroxisome function from the structure of pex3 in complex with a soluble fragment of pex19. *J. Biol. Chem.* **285,** 25410–25417 (2010).
19. Neufeld, C. *et al.* Structural basis for competitive interactions of pex14 with the import receptors pex5 and pex19. *EMBO J.* **28,** 745–754 (2009).
20. Fransen, M. *et al.* Analysis of human pex19p's domain structure by pentapeptide scanning mutagenesis. *J. Mol. Biol.* **346,** 1275–1286 (2005).
21. Schueller, N. *et al.* The peroxisomal receptor pex19p forms a helical mpts recognition domain. *EMBO J.* **29,** 2491–2500 (2010).
22. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372,** 774–797 (2007).
23. Sacksteder, K. A. *et al.* Pex19 binds multiple peroxisomal membrane proteins, is predominantly cytoplasmic, and is required for peroxisome membrane synthesis. *J. Cell Biol.* **148,** 931–944 (2000).
24. Halbach, A. *et al.* Function of the pex19-binding site of human adrenoleukodystrophy protein as targeting motif in man and yeast. Pmp targeting is evolutionarily conserved. *J. Biol. Chem.* **280,** 21176–21182 (2005).
25. Dominguez, C., Boelens, R. & Bonvin, A. M. Haddock: A protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125,** 1731–1737 (2003).
26. Chandra, A. *et al.* The gdi-like solubilizing factor pdedelta sustains the spatial organization and signalling of ras family proteins. *Nat. Cell Biol.* **14,** 148–158 (2012).
27. Ismail, S. A. *et al.* Arl2-gtp and arl3-gtp regulate a gdi-like transport system for farnesylated cargo. *Nat. Chem. Biol.* **7,** 942–949 (2011).
28. Lane, K. T. & Beese, L. S. Thematic review series: Lipid posttranslational modifications. Structural biology of protein farnesyltransferase and geranylgeranyltransferase type i. *J. Lipid Res.* **47,** 681–699 (2006).
29. Caplan, A. J., Tsai, J., Casey, P. J. & Douglas, M. G. Farnesylation of ydj1p is required for function at elevated growth temperatures in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **267,** 18890–18895 (1992).
30. Gotte, K. *et al.* Pex19p, a farnesylated protein essential for peroxisome biogenesis. *Mol. Cell. Biol.* **18,** 616–628 (1998).
31. Delaglio, F. *et al.* Nmrpipe: a multidimensional spectral processing system based on unix pipes. *J. Biomol. NMR* **6,** 277–293 (1995).
32. Goddard, T. & Kneller, D. Sparky 3 (University of California, San Francisco).
33. Jung, Y. S. & Zweckstetter, M. Mars—robust automatic backbone assignment of proteins. *J. Biomol. NMR* **30,** 11–23 (2004).
34. Sattler, M., Schleucher, J. & Griesinger, C. Heteronuclear multidimensional nmr experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nucl. Magn. Reson. Spectrosc.* **34,** 93–158 (1999).
35. Farrow, N. A. *et al.* Backbone dynamics of a free and phosphopeptide-complexed src homology 2 domain studied by 15n nmr relaxation. *Biochemistry* **33,** 5984–6003 (1994).
36. Madl, T., Bermel, W. & Zangger, K. Use of relaxation enhancements in a paramagnetic environment for the structure determination of proteins using nmr spectroscopy. *Angew. Chem. Int. Ed. Engl.* **48,** 8259–8262 (2009).
37. Hocking, H. G., Zangger, K. & Madl, T. Studying the structure and dynamics of biomolecules by using soluble paramagnetic probes. *Chemphyschem.* **14,** 3082–3094 (2013).
38. Gobl, C., Madl, T., Simon, B. & Sattler, M. Nmr approaches for structural analysis of multidomain proteins and complexes in solution. *Prog. Nucl. Magn. Reson. Spectrosc.* **80,** 26–63 (2014).
39. Guntert, P. Automated structure determination from nmr spectra. *Eur. Biophys. J.* **38,** 129–143 (2009).
40. Schuttelkopf, A. W. & van Aalten, D. M. Prodrg: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr. D Biol. Crystallogr.* **60,** 1355–1363 (2004).
41. Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. Talos+: a hybrid method for predicting protein backbone torsion angles from nmr chemical shifts. *J. Biomol. NMR* **44,** 213–223 (2009).
42. Brunger, A. T. *et al.* Crystallography & nmr system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54,** 905–921 (1998).
43. Huybrechts, S. J. *et al.* Identification of a novel PEX14 mutation in Zellweger syndrome. *J. Med. Genet.* **45,** 376–383 (2008).
44. Saveria, T. *et al.* Conservation of pex19-binding motifs required for protein targeting to mammalian peroxisomal and trypanosome glycosomal membranes. *Eukaryot. Cell* **6,** 1439–1449 (2007).
45. Will, G. K. *et al.* Identification and characterization of the human orthologue of yeast pex14p. *Mol. Cell. Biol.* **19,** 2265–2277 (1999).

Acknowledgements

We are grateful to A. Itzen and H. Kooshapur (TU München) for critical reading of the manuscript; S. Holton, N. Schüller, C. Williams, K. Fodor (EMBL Hamburg) for expression plasmids, material and discussions and N. Raschke (TU München), E. Becker and J. Wolf (University Bochum) for help with sample preparation and experiments. We are grateful to J. Boisbouvier, IBS Grenoble for donating the precursor for stereoselective methyl-labelling; S. Duhr and S. Blanke (NanoTemper Technologies GmbH, Munich) for support and H. Waldmann (MPI Dortmund) for spin label precursors. We thank the Bavarian NMR Centre (BNMRZ) for NMR measurement time. This work was supported by the Deutsche Forschungsgemeinschaft (SFB594, FOR1905 to M.S., ER178/4-1 to E.R.) and the EU FP7 NMI3 project (M.S.). U.S. acknowledges support from the Elitenetzwerk Bayern; K.T. from the Ministry of Education, Youth and Sports of the Czech Republic under the project CEITEC 2020 (LQ1601); T.M. from an EMBO Long Term Fellowships, a Schrödinger Fellowship by the Austrian Science Fund (FWF) and an APART-fellowship (Austrian Academy of Sciences) as well as the Bavarian Ministry of Sciences, Research and the Arts in the framework of the Bavarian Molecular Biosystems Research Network. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

The author(s) have made the following declarations about their contributions: U.S. and L.E. performed molecular biology, protein biochemistry, NMR data acquisition and analysis and biophysical binding studies, K.T. structure calculation and NMR experiments. T.M. contributed to NMR and performed solvent PRE calculations. R.R. performed hydrophobic interaction chromatography. W.S., R.E. and J.R. designed and conducted *in vivo* analyses. M.S., W.S., R.E., U.S., K.T., M.W. and L.E. designed the project and wrote the manuscript. All authors discussed the results and commented on the manuscript.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Emmanouilidis, L. *et al.* Allosteric modulation of peroxisomal membrane protein recognition by farnesylation of the peroxisomal import receptor PEX19. *Nat. Commun.* **8,** 14635 doi: 10.1038/ncomms14635 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2017

Journal of Biomolecular NMR (2019) 73:183–189
https://doi.org/10.1007/s10858-019-00241-9

ARTICLE

Selective isotope labeling for NMR structure determination of proteins in complex with unlabeled ligands

Konstantinos Tripsianes[1] · Ulrike Schütz[2,3] · Leonidas Emmanouilidis[2,3] · Gerd Gemmecker[2,3] · Michael Sattler[2,3]

Received: 10 December 2018 / Accepted: 19 March 2019 / Published online: 30 April 2019
© The Author(s) 2019

Abstract

The physiological role of proteins is frequently linked to interactions with non-protein ligands or posttranslational modifications. Structural characterization of these complexes or modified proteins by NMR may be difficult as the ligands are usually not available in an isotope-labeled form and NMR spectra may suffer from signal overlap. Here, we present an optimized approach that uses specific NMR isotope-labeling schemes for overcoming both hurdles. This approach enabled the high-resolution structure determination of the farnesylated C-terminal domain of the peroxisomal protein PEX19. The approach combines specific ^{13}C, ^{15}N and ^{2}H isotope labeling with tailored NMR experiments to (i) unambiguously identify the NMR frequencies and the stereochemistry of the unlabeled 15-carbon isoprenoid, (ii) resolve the NMR signals of protein methyl groups that contact the farnesyl moiety and (iii) enable the unambiguous assignment of a large number of protein-farnesyl NOEs. Protein deuteration was combined with selective isotope-labeling and protonation of amino acids and methyl groups to resolve ambiguities for key residues that contact the farnesyl group. Sidechain-labeling of leucines, isoleucines, methionines, and phenylalanines, reduced spectral overlap, facilitated assignments and yielded high quality NOE correlations to the unlabeled farnesyl. This approach was crucial to enable the first NMR structure of a farnesylated protein. The approach is readily applicable for NMR structural analysis of a wide range of protein–ligand complexes, where isotope-labeling of ligands is not well feasible.

Keywords NMR spectroscopy · Isotope labeling · Protein–ligand interactions · NOE

Introduction

Detailed understanding of the mechanism for protein–ligand interactions or the effects of posttranslational modifications requires structural information about the ligand-bound protein of interest on a molecular level. These complexes or modified proteins often possess intrinsic flexibility or low ligand affinity and are difficult to crystallize in the native conformation. Structural analysis by NMR on the other hand, can be problematic as isotope-labeling of the ligands may not be readily available, thus narrowing down the options to derive NOE-based distance restraints to ^{1}H only homonuclear experiments. However, homonuclear experiments suffer from spectral overlap, and thus render investigations of more complex systems involving larger proteins or chemically complex ligands challenging or unfeasible. On the other hand, various approaches for selective isotope enrichment of proteins based on metabolic percursors have been developed in the past decade and are mostly commercially available (Tugarinov et al. 2006; Kerfah et al. 2015).

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10858-019-00241-9) contains supplementary material, which is available to authorized users.

✉ Konstantinos Tripsianes
kostas.tripsianes@ceitec.muni.cz

✉ Michael Sattler
sattler@helmholtz-muenchen.de

[1] CEITEC—Central European Institute of Technology, Masaryk University, Kamenice 5, 62500 Brno, Czech Republic

[2] Institute of Structural Biology, Helmholtz Zentrum München, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany

[3] Center for Integrated Protein Science Munich at Chair of Biomolecular NMR, Department Chemie, Technische Universität München, Lichtenbergstr. 4, 85747 Garching, Germany

Protein prenylation, a class of lipid modification involving covalent addition of either farnesyl or geranylgeranyl groups, plays diverse roles in the regulation of trafficking, signaling and behavior of cellular proteins (Miura and Treisman 2006; Sorek et al. 2009; Novelli and D'Apice 2012). It is estimated that over 100 human proteins are potential substrates for farnesylation and geranylgeranylation (McTaggart 2006). Many of these proteins belong to the superfamily of small guanine nucleotide triphosphatases (GTPases) (Takai et al. 2001).

The peroxisomal membrane protein receptor PEX19 is farnesylated in vivo and the lipid modification is necessary for its critical function in peroxisomal biogenesis (Emmanouilidis et al. 2017; Rucktäschel et al. 2009). The 15-carbon isoprenoid, donated by farnesyl diphosphate (FPP), is transferred by protein farnesyl transferases to a conserved cysteine at the C-terminal CaaX motif (C = cysteine, aa = two aliphatic amino acids, X = any amino acid) (Manne et al. 1990). While the crystal structure of a C-terminal fragment of the PEX19 domain has been solved (Schueller et al. 2010), crystallization attempts with the farnesylated protein have failed. Therefore, solution state NMR techniques were applied for structural investigations of the farnesylated PEX19 C-terminal domain (residues 161–299), herein referred to as PEX19.

In this study, we made use of optimized isotope labeling of precursors in different protein labeling schemes. The combination of the various labeling protocols with individually adapted NMR experiments enabled us to circumvent the restrictions for the unlabeled ligand, to obtain complete resonance assignments for both protein and the attached lipid and to determine the first NMR structure of a farnesylated protein (Emmanouilidis et al. 2017) (Fig. 1).

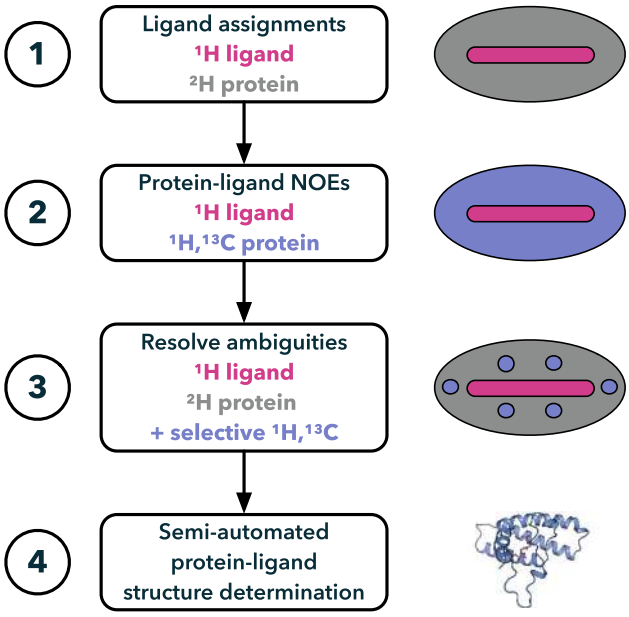


Fig. 1 Flowchart and NMR labeling schemes that allow identifying ligand frequencies (1), collecting protein–ligand NOEs (2), and resolving NOE ambiguities (3) for structure determination of protein–ligand complexes (4)

Results

NMR analysis of the farnesyl resonances

The farnesyl moiety is a linear grouping of three isoprene units that is attached enzymatically to the cysteine of the C-terminal CaaX box of PEX19. The reaction can be performed in vitro using purified PEX19, the farnesyl transferase enzyme, farnesyl pyrophosphate as the donor of (2*E*,6*E*)-farnesol, and cofactors such as Mg^{2+} and Zn^{2+} to obtain fully farnesylated PEX19 protein (Emmanouilidis et al. 2017) (Fig. 2a). All protein samples described herein refer to farnesylated PEX19. PEX19 labeling can vary but the farnesyl is always at natural isotopic abundance.

For characterizing the farnesyl group, we first used a PEX19 sample labeled uniformly with ^{15}N and ^{13}C, on which all the standard NMR experiments were performed. We applied isotope filters and recorded homonuclear NOESY and TOCSY spectra in D_2O (Breeze 2000; Sattler et al. 1999). By comparing the proton correlations in both spectra we could identify one isoprenoid group due to the very characteristic resonances of a pair of allylic methylene protons (Fig. S1). Inspection of the NOE correlations originating from the methylene protons of the lipidated cysteine in a 3D ^{13}C-edited NOESY spectrum (with or without isotope filters in ω_1) indicated that this is the first isoprenoid unit of the farnesyl (Fig. S1). Further farnesyl assignments were not possible due to signal overlap.

To identify the farnesyl resonances, we recorded a $^{1}H,^{13}C$ HSQC spectrum on a ^{13}C, ^{15}N, ^{2}H labeled PEX19 sample. However, in these spectra the natural abundance ^{13}C correlations of the farnesyl group are obscured by residual correlations from the protein (Fig. S2). According to the glucose specifications used (99% ^{13}C, 97% ^{2}H) and assuming that all other components of the bacterial growth medium were dissolved in D_2O, ca. 3% residual ^{1}H signal is expected for the protein. On the other hand, the attached farnesyl corresponds to 1% ^{1}H signal at natural isotopic abundance. Thus, residual protein and natural abundance farnesyl ^{13}C correlations are present at comparable intensities. Therefore, it is difficult to distinguish farnesyl correlations from the residual signals of the protein (Fig. S2).

To circumvent this problem we prepared perdeuterated PEX19 (97% ^{2}H) without ^{13}C labeling to reduce the occurrence of ^{13}C-bound protons in the protein (at natural abundance) to 0.03%. As shown in Fig. 2b, with this labeling

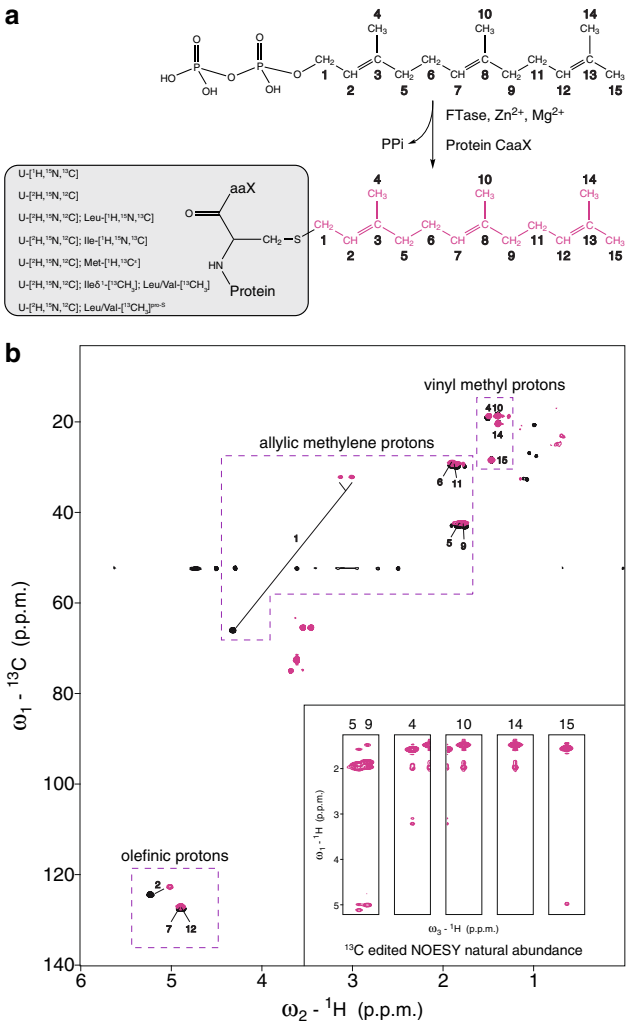


Fig. 2 Sampling the farnesyl resonances. **a** Schematic representation of in vitro farnesylation reaction for any target protein containing the CaaX box. Farnesyl groups are numbered. Any labeling scheme, as the ones listed for the present study, can be applied to the protein before the attachment of the unlabeled farnesyl. **b** Natural abundance ^{13}C,^{1}H HSQC spectra of farnesylated PEX19 labeled with [^{2}H]-glucose (magenta) and farnesyl pyrophosphate dissolved in methanol (black). In either case, farnesyl resonances are numbered as shown in **a**. Inset shows NOE strips from a natural abundance ^{13}C-edited 3D NOESY spectrum recorded on perdeuterated PEX19 using non-uniform sampling. Numbers at the top indicate the edited farnesyl resonances

scheme the farnesyl resonances are readily detected at natural abundance without interference from protein signals. By comparing the ^{1}H, ^{13}C correlations of the farnesyl with previous assignments (Umetsu et al. 1999), we could unambiguously identify all chemical shifts. In addition, we recorded a 3D ^{13}C-edited NOESY spectrum using non-uniform sampling (Hyberts et al. 2012) and a ^{13}C spectral window spanning the allylic methylene and vinyl methyl frequencies. The short-range NOE correlations observed in this experiment further confirmed the farnesyl chemical shift assignments (Fig. 2b). The ^{13}C frequencies of the vinyl methyl groups in combination with their NOE correlations established and confirmed the stereochemistry of (2*E*,6*E*)-farnesyl. The methyl groups of the terminal isoprenoid unit were stereospecifically assigned based on their ^{13}C chemical shifts. The *trans* methyl group at C14 has an upfield shifted carbon frequency that is in the same range as methyls C4 and C10 that are also in *trans* (~20 p.p.m.), and the *cis* methyl group (C15) has a downfield shifted carbon frequency (~28 p.p.m.). In support of the stereochemistry and our stereospecific assignments only the *cis* methyl group (C15) gives an NOE to the olefinic proton of the same isoprenoid unit (Fig. 2b). This correlation is not present for any of the *trans* methyl groups (C4, C10, and C14).

Finally, we compared the farnesyl fingerprint spectrum when covalently linked to PEX19 with that of the precursor, farnesyl pyrophosphate in methanol solution. Surprisingly, the chemical shifts are very similar (Fig. 2b). The only differences involve the allylic and olefinic atoms of the first isoprenoid unit and reflect the local microenvironment induced by the neighboring phosphates in pyrophosphate or the protein. The large carbon difference of the first allylic group is due to the different heteroatom attached to when in pyrophosphate or in cysteine. The other chemical shifts are highly comparable, suggesting that the farnesyl moiety experiences a protein hydrophobic environment similar to the organic solvent (Fig. 2b).

Protein-farnesyl NOE correlations

To obtain protein-farnesyl NOEs we initially recorded isotope-filtered and -edited NOESY experiments with a sample of PEX19 uniformly labeled with ^{15}N and ^{13}C. These experiments indicated a large number of NOE contacts between protein methyl groups and farnesyl. However, the analysis was complicated by limited spectral resolution in the ^{13}C dimension and severe overlap for some of the protein methyl frequencies. PEX19 contains a large number of aliphatic amino acids, including 17 leucine, 9 methionine, and 6 isoleucine residues, for which we could not obtain complete and unambiguous chemical shift assignments using conventional NMR methods and inspection of NOESY spectra (Fig. S3).

To complete chemical shift assignments and resolve the remaining ambiguities we prepared perdeuterated PEX19 with protonation of specific amino acid side chains or methyl groups (Metzler et al. 1996; Tugarinov et al. 2004). For leucine and isoleucine uniformly ^{15}N/^{13}C labeled amino acid was used, whereas for methionine a precursor with specific ^{13}C labeling of the ε carbon was employed. As shown in Fig. 3, amino acid selective labeling reduces spectral overlap and enabled unambiguous analysis of each residue. Two important aspects of the amino acid selective labeling

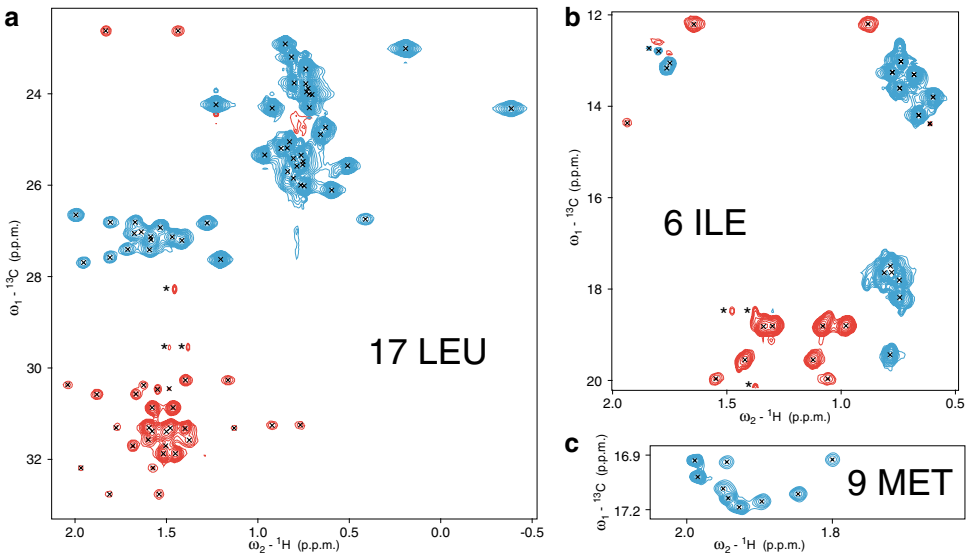


Fig. 3 NMR spectra recorded for amino acid selectively-labeled, perdeuterated and uniformly ^{15}N-labeled PEX19 protein with unlabeled farnesyl covalently attached. **a** Constant time ^{13}C,^{1}H HSQC spectrum of perdeuterated PEX19 expressed with ^{1}H/^{15}N/^{13}C leucine. **b** Constant time ^{13}C,^{1}H HSQC spectrum of perdeuterated PEX19 expressed with ^{1}H/^{15}N/^{13}C isoleucine. **c** ^{13}C,^{1}H HSQC spectrum of perdeuterated PEX19 expressed with ^{1}H/^{13}Cε methionine. In all cases the Cα proton of the amino acid precursor has been replaced by a deuteron of the solvent. All ^{1}H–^{13}C resonance pairs are observed. Positive peaks are coloured cyan and negative peaks are coloured red. Asterisks indicate natural abundance methyl crosspeaks of the attached farnesyl

employed is that no ^{13}C scrambling occurs for Leu, Ile, and Met as they are end products of their metabolic pathway (Lacabanne et al. 2017). Also, Hα protons are replaced by deuterons during protein synthesis, as reported previously (Metzler et al. 1996; Crespi et al. 1968; Crespi and Katz 1969; LeMaster 1989) (Fig. S4).

We compared the intensities of Cα–Hα and Cγ–Hγ correlations in a ^{1}H, ^{13}C HSQC spectrum of the leucine-labeled sample and estimated the level of deuteration at the α position to ~97%, consistent with the deuterium enrichment in the growth medium. In ^{13}C-edited 3D NOESY experiments all carbon frequencies are resolved, especially those of the methyl groups (Fig. 3). Although no Hα correlations are observable, we were able to assign all leucine ^{1}H and ^{13}C signals and confirm the isoleucine and methionine assignments based on intraresidue NOEs.

A large number of NOEs are observed between methionine residues and olefinic as well as vinyl methyl groups of the farnesyl moiety. Allylic methylene proton signals overlap with the diagonal peaks of the methionine methyl group (Fig. 4a), and thus NOEs could not be assigned. In some favorable cases farnesyl NOEs from leucine and isoleucine spectra were resolved as well, but most cross peaks are masked by the stronger intraresidue NOE correlations.

To obtain NOE correlations to aromatic protein signals we prepared a perdeuterated protein sample containing protonated phenylalanine residues (Vuister et al. 1994). The aromatic proton frequencies of the four phenylalanines are well resolved from the farnesyl ones, and a 2D homonuclear NOESY spectrum recorded in D_2O yielded NOEs to farnesyl from one of the aromatic rings (Fig. 4a). Since there is no degradation pathway for Phe in *E. coli*, no scrambling was observed for the aromatic sidechain (Lacabanne et al. 2017), but similarly to Leu and Ile the α protons were replaced by deuterons of the medium during bacterial protein production.

Using this strategy complete and unambiguous methyl assignments were obtained for PEX19. We next prepared an ILV-labeled PEX19 sample by adding α-ketobutyrate and α-ketoisovalerate in the growth media for selective methyl protonation of valine, leucine and δ1 group of isoleucine (Tugarinov et al. 2004). As shown in Fig. 4, these protein frequencies are distinct from the farnesyl ones. Therefore, a large number of protein-farnesyl NOE correlations could be unambiguously assigned, which would otherwise be masked by protein–protein correlations. Finally, an acetolactate precursor was used to obtain stereospecific assignments of methyl groups for leucine and valine residues (Gans et al. 2010) (Fig. 4b).

Structure calculation

Chemical shift assignments for the 15.5 kDa farnesylated PEX19 protein were complete to 99%. All protein heteronuclear spin pairs were resolved in NOESY spectra, and for Leu and Ile residues in spectra of selectively labeled

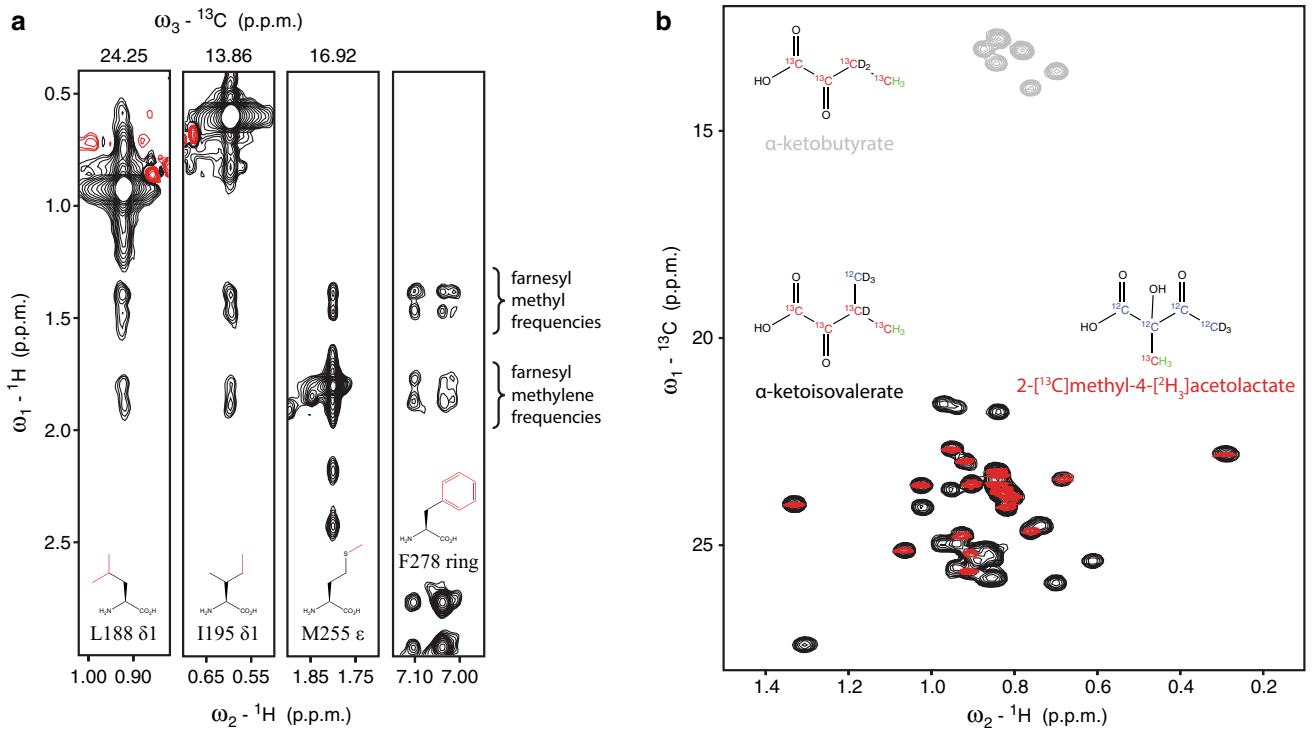


Fig. 4 Labeling schemes for mapping protein-farnesyl contacts. **a** Strips from ^{13}C-edited NOESY spectra recorded on samples specifically labeled for the methyl groups of leucine, isoleucine, and methionine (highlighted in red in the schematic representation of amino acids) and the aromatic portion of a 2D homonuclear NOESY spectrum on a PEX19 sample reversed labeled for phenylalanines. **b** ^{13}C, ^{1}H HSQC spectra recorded on ILV (U-[^{2}H,^{12}C]; Ileδ^{1}-[^{13}CH$_3$] gray; Leu/Val-[^{13}CH$_3$] black) and on (U-[^{2}H,^{12}C]; Leu/Val-[^{13}CH$_3$]$^{pro-S}$ red) PEX19 samples

samples described above. The proton resonances of the farnesyl group were correctly identified, but their chemical shift degeneracy, in particular between equivalent positions in the second and third isoprenoid units, resulted in ambiguity as to which farnesyl atoms of the second and third isoprenoid units are involved in a given NOE. Only the frequencies of the first isoprenoid unit (C1, C2, and C4) and the *cis* methyl group of the third isoprenoid unit (C15) could be resolved (for nomenclature see Fig. 2) in the 3D NOESY spectra. NMR signals corresponding to these farnesyl atoms were manually assigned in all spectra. Signals intensities of the NOE correlation cross peaks were automatically calibrated and converted to distance restraints with an optimized average-distance-parameter of CYANA (Guntert 2009). In a first round of structure calculations, NOE assignments of cross-peaks involving these farnesyl atoms were fixed and the remaining farnesyl NOEs were treated as ambiguous. By reference to the preliminary three-dimensional structure we could uniquely assign cross-peaks to the methyl (C10 and C14) and olefinic atoms (C7 and C12) of the second and third isoprenoid unit that have identical proton chemical shifts. In an iterative procedure additional peaks were held constant to account for distance restraints involving these farnesyl atoms. The NOE correlations to the methylene atoms of the second and third isoprenoid unit (C5, C6, C9, and C11) were assigned last in an automated fashion by using a stringent chemical shift tolerance of 0.015 ppm. Only two of these peaks were excluded by the program due to large violations of the corresponding distance restraints in the resulting structures. Both unassigned peaks matched methylene frequencies but they were singular and just above the noise level. On the contrary, all methylene correlations assigned by the program were present as pairs and corresponded to sequential methylene atoms (C5 and C6 or C9 and C11) (Fig. 4a).

Table 1 lists the peak assignment completeness for data recorded in order to map selectively the PEX19-farnesyl contacts. None of the original NOE cross peaks that were picked above a certain threshold were pruned during the iterative assignment strategy. By using a step-by-step semi-automated approach, we were able to collect 203 distance restraints between PEX19 and farnesyl, amounting on average to 17 distance restraints per farnesyl atom. The obtained NOEs are highlighted onto the final structure in Fig. 5.

These data were sufficient for determining a high-resolution structure that revealed a bent conformation of the lipid moiety buried inside a hydrophobic cavity of the protein (PDB:5LNF) (Emmanouilidis et al. 2017).

Table 1 Protein samples for collecting NOEs to farnesyl

Precursor (with perdeuterated PEX19 and unlabeled farnesyl)	NOESY spectra[a]	Peaks[b]	Assigned (%)
$^{1}H/^{15}N/^{13}C$ leucine[c]	^{13}C-edited 3D NOESY	390	384 (98%)
$^{1}H/^{15}N/^{13}C$ isoleucine[c]	^{13}C-edited 3D NOESY	129	127 (98%)
$^{1}H/^{13}C\varepsilon$ methionine	^{13}C-edited 3D NOESY	34	34 (100%)
Phenylalanine	2D NOESY	26	26 (100%)
3-[$^{2}H_3$]methyl-3-[^{2}H]-1,2,3,4-[^{13}C]-α-ketoisovalerate and 3-[$^{2}H_2$]-1,2,3,4-[^{13}C]-α-ketobutyrate	^{13}C-edited 3D NOESY	96	94 (98%)
2-[^{13}C] methyl-4-[$^{2}H_3$]-acetolactate[d]	–	–	–

[a]Leucine, isoleucine and a-ketoacids were recorded in a constant-time fashion

[b]Each set of peaks contains protein–protein NOEs in addition to protein-farnesyl NOEs

[c]Unassigned peaks involve NOEs to protein amide protons

[d]Due to low sample concentration 3D NOESY spectra were not feasible

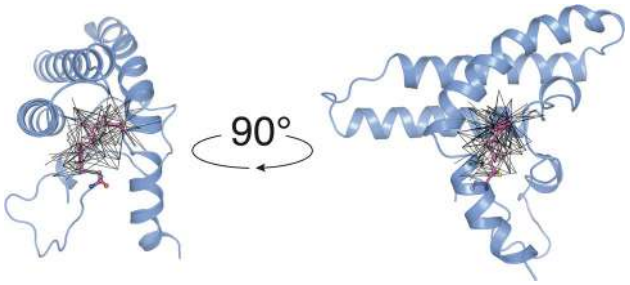


Fig. 5 Two views of the solution NMR structure of farnesylated PEX19 (PDB:5LNF). Each black line corresponds to an NOE distance restrain between the protein and the farnesyl moiety. The distance network comprises 203 NOEs recorded on samples using various labeling schemes

Conclusion

We report an optimized protocol for NMR structure calculation of proteins, where ligands or covalently attached modifications cannot be readily obtained in isotope-labeled form. The approach combines specific ^{13}C, ^{15}N and ^{2}H isotope labeling with tailored NMR experiments for chemical shift and NOE assignments to overcome problems with signal overlap and protein–ligand NOE ambiguities. The approach is generally applicable to other challenging systems, where chemical shift degeneracy and signal overlap render analysis of NOESY spectra difficult and prohibit convergence of structure calculations. Of course, the overall success depends on additional factors unique to the system under investigation, e.g. strength of the interaction, exchange processes induced by binding, number of ligand chemical shifts and their degeneracy. However, the basic principles presented here should enable to assess the complexity of the system, adapt the individual steps accordingly and supplement the workflow with additional experiments, if required.

Methods

Protein expression and isotope labeling

PEX19 was expressed from a pETM-11 vector containing an N-terminus His-6 tag followed by a TEV cleavage site. *Escherichia coli* BL21 (DE3) cells were grown at 37 °C in perdeuterated medium supplied with U-[^{2}H]-D-glucose (2 g l^{-1}) and $^{15}NH_4Cl$ (1 g l^{-1}) as the sole carbon and nitrogen sources, respectively. When OD_{600} reached ~ 0.5, 50 mg of the respective labeled amino acid or 150 mg of the respective precursor was added. Cultures were kept at 37 °C for another hour, then incubated at 20 °C for 15 min and induced with 0.5 mM IPTG. After overnight expression at 20 °C, cells were harvested by centrifugation at 5000 r.p.m., washed with PBS buffer and stored at − 80 °C.

Protein purification

Cell pellets were resuspended in 30 ml lysis buffer containing 50 mM HEPES (pH 7.5), 300 mM NaCl, 10 mM imidazole, 1 mM tris-(2-carboxyethyl)-phosphine (TCEP), and protease inhibitor mix, and lysed by sonication. After centrifugation at 14,000 g the supernatant was loaded onto Ni–NTA gravity-flow resin (Qiagen) and proteins were eluted with increasing concentrations of imidazole. TEV protease treatment was performed overnight (approximately 250 µg TEV for 10 mg of PEX19) and the His-6 tag was removed by Ni–NTA affinity.

In vitro farnesylation

Untagged PEX19 was mixed with farnesyl transferase fused to His-6 tag and farnesyl pyrophosphate (Sigma Aldrich) in a buffer containing the appropriate amounts of Zn^{2+} and Mg^{2+} ions (Caplan et al. 1992). The reaction mixture was incubated at 37 °C for 1 h and passed through Ni–NTA to

remove the farnesyl transferase. Farnesylated PEX19 was further purified by size exclusion chromatography and stored in NMR buffer (20 mM potassium phosphate, pH 6.5 and 50 mM NaCl). Farnesylation was confirmed by SDS gel electrophoresis, ^{1}H,^{15}N HSQC NMR spectra, and mass spectrometry (Emmanouilidis et al. 2017).

NMR spectroscopy

NMR spectra were acquired at 298 K on an Avance 900 instrument equipped with a TCI cryo-probe. Spectra were processed using NMRPipe (Delaglio et al. 1995) and analyzed with Sparky (Goddard and Kneller 1996).

Acknowledgements This research was financially supported by the Deutsche Forschungsgemeinschaft [FOR1905 PERTRANS and GRK1721 to M.S.], and the Ministry of Education, Youth and Sports of Czech Republic within the CEITEC 2020 project [LQ1601 to K.T.]. We acknowledge the Bavarian NMR Centre (BNMRZ) for NMR measurement time.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

Breeze AL (2000) Isotope-filtered NMR methods for the study of biomolecular structure and interactions. Prog Nucl Magn Reson Spectrosc 36:323–372

Caplan AJ, Tsai J, Casey PJ, Douglas MG (1992) Farnesylation of YDJ1p is required for function at elevated growth temperatures in Saccharomyces cerevisiae. J Biol Chem 267:18890–18895

Crespi HL, Katz JJ (1969) High resolution proton magnetic resonance studies of fully deuterated and isotope hybrid proteins. Nature 224:560–562

Crespi HL, Rosenberg RM, Katz JJ (1968) Proton magnetic resonance of proteins fully deuterated except for 1H-leucine side chains. Science 161:795–796

Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 6:277–293

Emmanouilidis L, Schutz U, Tripsianes K, Madl T, Radke J, Rucktaschel R, Wilmanns M, Schliebs W, Erdmann R, Sattler M (2017) Allosteric modulation of peroxisomal membrane protein recognition by farnesylation of the peroxisomal import receptor PEX19. Nat Commun 8:14635

Gans P, Hamelin O, Sounier R, Ayala I, Dura MA, Amero CD, Noirclerc-Savoye M, Franzetti B, Plevin MJ, Boisbouvier J (2010) Stereospecific isotopic labeling of methyl groups for NMR spectroscopic studies of high-molecular-weight proteins. Angew Chem Int Ed Engl 49:1958–1962

Goddard TD, Kneller DG (1996) Sparky 3

Guntert P (2009) Automated structure determination from NMR spectra. Eur Biophys J 38:129–143

Hyberts SG, Arthanari H, Wagner G (2012) Applications of non-uniform sampling and processing. Top Curr Chem 316:125–148

Kerfah R, Plevin MJ, Sounier R, Gans P, Boisbouvier J (2015) Methyl-specific isotopic labeling: a molecular tool box for solution NMR studies of large proteins. Curr Opin Struct Biol 32:113–122

Lacabanne D, Meier BH, Bockmann A (2017) Selective labeling and unlabeling strategies in protein solid-state NMR spectroscopy. J Biomol NMR

LeMaster DM (1989) Deuteration in protein proton magnetic resonance. Methods Enzymol 177:23–43

Manne V, Roberts D, Tobin A, O'Rourke E, De Virgilio M, Meyers C, Ahmed N, Kurz B, Resh M, Kung HF et al (1990) Identification and preliminary characterization of protein-cysteine farnesyltransferase. Proc Natl Acad Sci USA 87:7541–7545

McTaggart SJ (2006) Isoprenylated proteins. Cell Mol Life Sci 63:255–267

Metzler WJ, Wittekind M, Goldfarb V, Mueller L, Farmer BT (1996) Incorporation of $^{1}H/^{13}C/^{15}N$-{Ile, Leu, Val} into a perdeuterated, ^{15}N-labeled protein: potential in structure determination of large proteins by NMR. J Am Chem Soc 118:6800–6801

Miura GI, Treisman JE (2006) Lipid modification of secreted signaling proteins. Cell Cycle 5:1184–1188

Novelli G, D'Apice MR (2012) Protein farnesylation and disease. J Inherit Metab Dis 35:917–926

Rucktäschel R, Thoms S, Sidorovitch V, Halbach A, Pechlivanis M, Volkmer R, Alexandrov K, Kuhlmann J, Rottensteiner H, Erdmann R (2009) Farnesylation of pex19p is required for its structural integrity and function in peroxisome biogenesis. J Biol Chem 284:20885–20896

Sattler M, Schleuchter J, Griesinger C (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. Prog NMR Spectrosc 34:93–158

Schueller N, Holton SJ, Fodor K, Milewski M, Konarev P, Stanley WA, Wolf J, Erdmann R, Schliebs W, Song YH, Wilmanns M (2010) The peroxisomal receptor Pex19p forms a helical mPTS recognition domain. EMBO J 29:2491–2500

Sorek N, Bloch D, Yalovsky S (2009) Protein lipid modifications in signaling and subcellular targeting. Curr Opin Plant Biol 12:714–720

Takai Y, Sasaki T, Matozaki T (2001) Small GTP-binding proteins. Physiol Rev 81:153–208

Tugarinov V, Hwang PM, Kay LE (2004) Nuclear magnetic resonance spectroscopy of high-molecular-weight proteins. Annu Rev Biochem 73:107–146

Tugarinov V, Kanelis V, Kay LE (2006) Isotope labeling strategies for the study of high-molecular-weight proteins by solution NMR spectroscopy. Nat Protoc 1:749–754

Umetsu M, Wang ZY, Kobayashi M, Nozawa T (1999) Interaction of photosynthetic pigments with various organic solvents. Magnetic circular dichroism approach and application to chlorosomes. Biochim Biophys Acta 1410:19–31

Vuister GW, Kim S-J, Wu C, Bax A (1994) 2D and 3D NMR study of phenylalanine residues in proteins by reverse isotopic labeling. J Am Chem Soc 116:9206–9210

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Hanáková et al. Cell Communication and Signaling (2019) 17:170
https://doi.org/10.1186/s12964-019-0470-z

Cell Communication and Signaling

RESEARCH Open Access

Comparative phosphorylation map of Dishevelled 3 links phospho-signatures to biological outputs

Kateřina Hanáková[1,2†], Ondřej Bernatík[3,4†], Marek Kravec[3], Miroslav Micka[3], Jitender Kumar[1], Jakub Harnoš[3], Petra Ovesná[5], Petra Paclíková[3], Matěj Rádsetoulal[3], David Potěšil[1], Konstantinos Tripsianes[1], Lukáš Čajánek[4], Zbyněk Zdráhal[1,2*] and Vítězslav Bryja[3,6*]

Abstract

Background: Dishevelled (DVL) is an essential component of the Wnt signaling cascades. Function of DVL is controlled by phosphorylation but the molecular details are missing. DVL3 contains 131 serines and threonines whose phosphorylation generates complex barcodes underlying diverse DVL3 functions. In order to dissect the role of DVL phosphorylation we analyzed the phosphorylation of human DVL3 induced by previously reported (CK1ε, NEK2, PLK1, CK2α, RIPK4, PKCδ) and newly identified (TTBK2, Aurora A) DVL kinases.

Methods: Shotgun proteomics including TiO_2 enrichment of phosphorylated peptides followed by liquid chromatography tandem mass spectrometry on immunoprecipitates from HEK293T cells was used to identify and quantify phosphorylation of DVL3 protein induced by 8 kinases. Functional characterization was performed by in-cell analysis of phospho-mimicking/non-phosphorylatable DVL3 mutants and supported by FRET assays and NMR spectroscopy.

Results: We used quantitative mass spectrometry and calculated site occupancies and quantified phosphorylation of > 80 residues. Functional validation demonstrated the importance of CK1ε-induced phosphorylation of S268 and S311 for Wnt-3a-induced β-catenin activation. S630–643 cluster phosphorylation by CK1, NEK2 or TTBK2 is essential for even subcellular distribution of DVL3 when induced by CK1 and TTBK2 but not by NEK2. Further investigation showed that NEK2 utilizes a different mechanism to promote even localization of DVL3. NEK2 triggered phosphorylation of PDZ domain at S263 and S280 prevents binding of DVL C-terminus to PDZ and promotes an open conformation of DVL3 that is more prone to even subcellular localization.

Conclusions: We identify unique phosphorylation barcodes associated with DVL function. Our data provide an example of functional synergy between phosphorylation in structured domains and unstructured IDRs that together dictate the biological outcome.

Keywords: Dishevelled, DVL3, Phosphorylation, Kinase, Mass spectrometry, CK1, TTBK2, NEK2, Wnt

Background

Wnt signaling pathway has been linked to an etiology of multiple developmental defects, inherited diseases and many types of cancer [1]. Wnt pathways can be divided into several main "branches". The best studied (canonical) pathway depends on β-catenin and members of the TCF/LEF (T-cell factor/lymphoid enhancer-binding factor) family of transcriptional factors. In the absence of Wnt ligand, the intracellular level of free β-catenin is constantly low due to the activity of a degradation complex including adenomatous polyposis coli, Axin and glycogen synthase kinase (GSK) 3β. Upon Wnt binding, the destruction complex is inhibited allowing accumulation and nuclear activity of β-catenin. However, Wnts can also activate other, so-called non-canonical Wnt pathways, which are β-catenin-independent and biochemically distinct from canonical Wnt signaling. There is strong

* Correspondence: zdrahal@sci.muni.cz; bryja@sci.muni.cz
†Kateřina Hanáková and Ondřej Bernatík contributed equally to this work.
[1]CEITEC—Central European Institute of Technology, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic
[3]Department of Experimental Biology, Faculty of Science, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic
Full list of author information is available at the end of the article

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

evidence that several such pathways exist (for complete overview see [2]).

All known Wnt-induced pathways are transduced through Frizzled transmembrane receptors [3] and intracellular protein Dishevelled (Dsh in *Drosophila*, DVL1–3 in human). DVL proteins consist of three structured domains: N-terminally located DIX domain, centrally located PDZ domain and C-terminally located DEP domain. The individual domains are linked by intrinsically disordered regions followed by approximately 200 amino acid long, also intrinsically disordered, C-terminal tail.

There is a general agreement based on genetic experiments that DVL plays a crucial role as a signaling hub in both Wnt/β-catenin and non-canonical Wnt pathways [4]. In addition DVL has been reported to have multiple other functions such as docking of basal body [5, 6], function and maintenance of primary cilia [7], cytokinesis [8], or positioning of the mitotic spindle [9]; all of them possibly linked to its role in the regulation of centrosomal cycle [10].

Despite the well documented role of DVL in the Wnt signaling and the growing evidence for its participation in additional cellular processes, the molecular mechanisms that regulate DVL action in the cell are almost unknown. However, it is believed that post-translational modifications (PTM), particularly phosphorylation, represent a key component of such regulatory mechanisms. DVL proteins are very rich in serine (S) and threonine (T); for example, DVL3 contains 131 S/T residues, which can be potentially phosphorylated. The best described consequence of the activation of both Wnt/β-catenin as well as non-canonical Wnt pathway is phosphorylation of DVL by the Wnt-induced Casein kinase 1 ε (CK1ε) [11–14]. In addition to CK1ε, multiple other kinases have been reported to phosphorylate DVL in different contexts. For example, CK2α in both Wnt signaling pathways [14–16], Polo-like kinase (PLK) 1 in the control of mitotic spindle [9], Nima-related kinase (NEK) 2 in the centrosome [10, 17], protein kinase C (PKC) δ in non-canonical Wnt signaling [18], and receptor-interacting protein kinase (RIPK) 4 in the Wnt/β-catenin signaling [19]. Recent work using mass spectrometry [10, 20, 21] has identified more than 50 S/T of DVL that are indeed phosphorylated. However, the functional significance of these phosphorylation sites in DVL remains unclear. With respect to the current understanding of PTMs in the intrinsically disordered proteins, it is reasonable to speculate that the physiological function is achieved by a combination ("barcode") of phosphorylated sites rather than by mere phosphorylation of individual sites.

In this study we analyzed and compared the phosphorylation of DVL3 induced by eight S/T kinases. We have selected DVL3 as a candidate because its absence shows the strong phenotypes in vivo [22]. We have studied the phosphorylation of this DVL isoform earlier [10, 16, 23, 24], which allows for a direct comparison. The studied kinases included six previously reported DVL kinases, Aurora A that was reported with DVL in the same complex [25] and TTBK2, a DVL kinase identified in this study. We have applied complementary proteomic approaches to describe in detail and in a quantitative manner how individual kinases modify DVL3. The phosphorylation barcoding unveiled unique but also common phosphorylation patterns that represent a reference point for the interpretation of existing and any future data analyzing phosphorylation of DVL. Last but not least, our work provides an example of universal pipelines for the phosphoanalysis of proteins modified in complex patterns at dozens of residues.

Materials and methods

Cell culture, transfection, and treatments

HEK293T, HEK293 T-REx and DVL1/2/3-null HEK293 T-REx cells were propagated in DMEM, 10% FCS, 2 mM L-glutamine, 50 units/ml penicillin, 50 units/ml streptomycin. CK1ε inhibitor PF-670462 was used at 10 μM. Cells were seeded on appropriate culture dishes (15 cm diameter dish for IP, 24-well plate for WB, dual luciferase assay, ICC) and the next day were transfected using polyethylenimine (PEI) in a stoichiometry of 3 μl PEI (0.1% w/v in MQ water) per 1 μg of DNA. Cells were harvested for immunoblotting or immunocytofluorescence 24 h after transfection, for immunoprecipitation after 48 h. The following plasmids have been published previously: FLAG-DVL3 [26], CK1ε [27], Myc-NEK2 [28], FLAG-PLK1 [29], Myc-Aurora A [30], HA-CK2α [31], YFP-PKCδ [32], VSV-RIPK4 [33] and GFP-TTBK2 WT and KD [34].

Dual luciferase TopFlash/Renilla reporter assay

For the luciferase reporter assay, cells were transfected with 0.1 μg of Super8X TopFlash construct, 0.1 μg of pRLTKluc (Renilla) luciferase construct and other plasmids as indicated (50 ng of DVL3 in Fig. 1d, 200 ng of DVL3 in Fig. 9b) in a 24-well plate and processed 24 h after transfection. For the TopFlash assay, a Promega dual luciferase assay kit was used according to the manufacturer's instructions. Relative luciferase units of firefly luciferase were measured and normalized to the Renilla luciferase signal.

Coimmunoprecipitation and Western blotting

For MS/MS-based identification of phosphorylation, HEK293T cells were seeded on 15 cm dishes and transfected with corresponding plasmids (8 μg of DVL3 plasmid plus 8 μg of pCDNA3 or plasmid encoding the kinase per dish) 24 h after seeding. Two ml of ice-cold lysis buffer supplemented with protease inhibitors

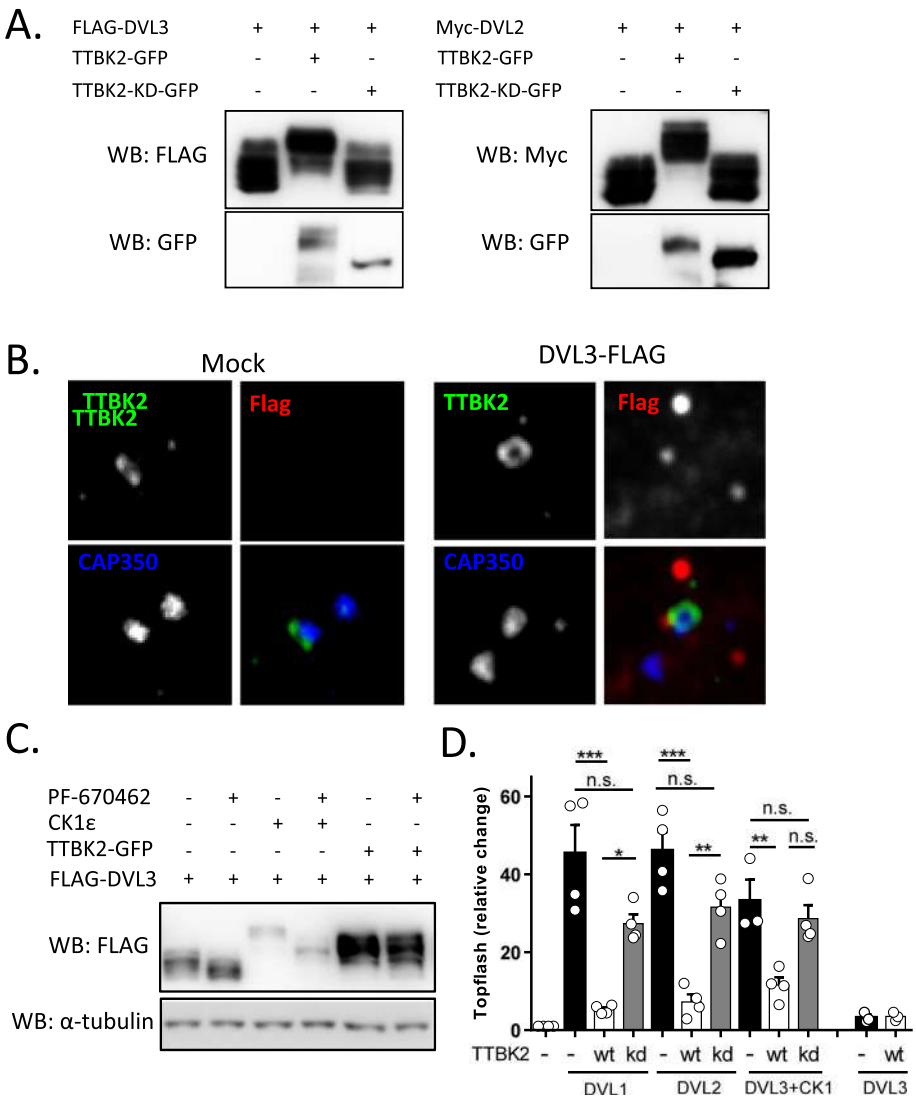


Fig. 1 Identification of TTBK2 as a novel DVL kinase. **a**: HEK293 cells were transfected with FLAG-DVL3 and Myc-DVL2 plasmids with wild type (wt) or kinase dead (KD) TTBK2-GFP. Active TTBK2 promoted phosphorylation-dependent mobility shift of DVL3 on Western blotting. **b**: Endogenous TTBK2 (green) localized into distal appendages of the mother centriole in hTERT-RPE1 cells (left). Overexpression of FLAG-DVL3 (stained in red) was not able to displace TTBK2 from the centriole (right). Centrioles were stained with CAP350 (blue). **c**: HEK293 cells were transfected with indicated plasmids, treated with CK1ε inhibitor PF-670462 (10 μM) and subsequently analyzed by Western blotting. TTBK2-induced electrophoretic mobility shift of DVL3 was not diminished upon CK1ε inhibition unlike the mobility shift induced by CK1ε. **d**: HEK293 cells were transfected with indicated plasmids and by the TopFLASH reporter system. Luminescence in the cell lysates was measured 24 h after transfection. Mean, SD and individual data points are indicated. Statistical differences were tested by One-way ANOVA and Tukey's post test (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$)

(Roche Applied Science, 11,836,145,001), phosphatase inhibitors (Calbiochem, 524,625), 0.1 mM DTT, and 10 mM N-ethylmaleimide (Sigma E3876) was used for lysis of one 15 cm dish. Lysate was collected after 20 min of lysis on 4 °C and was cleared by centrifugation at 18000×g for 20 min. Three μg of anti-FLAG M2 (F1804; Sigma) antibody were used per sample. Samples were incubated with the antibody for 40 min, then 25 μl of G protein-Sepharose beads (GE Healthcare, 17–0618-05) equilibrated in the lysis buffer were added to each sample. Samples were incubated on the carousel overnight, washed 6 times with lysis buffer and finally, 40 μl of 2× Laemmli buffer was added and samples were subsequently boiled.

The samples were loaded to 8% SDS-PAGE gels. Electrophoresis was preformed through the stacking gel at 100 V and through the separating gel at 150 with PageRuler Prestained Protein Ladder (Thermofisher, Cat. No. 26620) as a marker. Gels were then process for mass spectrometry or the proteins were transferred to the polyvinylidene difluoride membrane (Merck Millipore, Cat. No. IPVH00010) by Western Blotting (WB). Subsequently, the membrane was blocked

in 5% defatted milk or 3% BSA for 1 h with shaking. After the blocking step, the membrane was incubated with corresponding primary and secondary antibodies. Proteins were visualized by ECL (Enhanced chemiluminescence; Merck Millipore, Cat. No. WBKLS0500). The signal was detected using Vilber FUSION-SL system. Western blot was quantified using ImageJ software.

The antibodies used were: anti-FLAG M2 (Sigma-Aldrich, #F1804) for WB and IP, anti-CK1ε (Santa Cruz, #sc-6471), anti-GFP (Fitzgerald, #20R-GR-011), HA11 (Covance #MMS-101R), VSV (Sigma-Aldrich #V 5507), c-Myc (Santa Cruz, #sc-40), α-tubulin (Proteintech, 66, 031–1-Ig), anti-DVL3 (sc-8027, Santa Cruz Biotechnology) and anti-β-actin (cs4970, Cell Signalling). Following phospho-specific antibodies have been published previously – pS280-DVL3 (pS280) [10], pS643 [20], and pS697 [10]. Anti phospho-S192 antibody was prepared by immunizing rabbits by TTSFFDS(p) DEDDST peptide on a service basis by Moravian Biotechnology (http://www.moravian-biotech.com).

Immunofluorescence

Cells were seeded onto glass coverslips and next day transfected according to the scheme. 24 h post transfection medium was removed, cells were washed by PBS and fixed by 4% PFA for 10 min or ice-cold MetOH (5 min/− 20 °C, Fig. 1b). Coverslips were then washed in PBS and incubated with primary antibodies for 1 h, washed three times with PBS and then incubated with secondary antibodies conjugated to Alexa Fluor 488 (Invitrogen A11001) or/and Alexa Fluor 594 (Invitrogen A11058), washed with PBS and stained with DAPI (1:5000); all coverslips were mounted on microscopic slides. Cells were then visualized on Olympus IX51 fluorescent microscope using 40× air or 100× oil objectives and/or Olympus Fluoview 500 confocal laser scanning microscope IX71 using 100× oil objective. 200 positive cells per experiment ($N = 3$) were analyzed and scored according to their phenotype into two categories (punctae/even). The antibodies used were as follows: anti-FLAG M2 (Sigma-Aldrich, #F1804), anti-DVL3 (Santa Cruz, #sc-8027), anti-CK1ε (Santa Cruz, #sc-6471), HA11 (Covance #MMS-101R), VSV (Sigma-Aldrich #V 5507), c-Myc (Santa Cruz, #sc-40) and anti-GFP (Fitzgerald, #20R-GR-011), CAP350 [35], TTBK2 (Sigma-Aldrich, #HPA018113). Images presented in Fig. 1b were acquired using DeltaVision-Elite system (Applied Precision/GE) with a 100×/1.4 Apo plan oil immersion objective. Image stacks were taken with a z distance of 0.2 μm, deconvolved (conservative ratio, three cycles), and projected as maximal intensity images by using SoftWoRX (Applied Precision/GE).

FlAsH FRET

The HEK293 cells were seeded onto round 24-mm cover slips, which were previously placed in six-well plates and coated with 200 μl of poly-D-lysine (1 mg/ml) for 20 min. Cells were transfected 16–18 h later using Effectene (Qiagen), according to the manufacturer's instructions. DMEM was replaced 6 h later and the analysis was done 24 h after transfection.

FlAsH labeling of the DVL3 FlAsH III sensor was performed as previously described [24]. Shortly, transfected cells were washed once with Hank's Balanced Salt Solution (HBSS) containing 1.8 g/l glucose and then incubated at 37 °C for 1 h with HBSS supplemented with 500 nM FlAsH; 12.5 μM 1,2-ethanedithiol (EDT). In order to reduce non-specific labeling, cells were then rinsed once with HBSS and incubated for 10 min with HBSS containing 250 μM EDT and corresponding inhibitors. Finally, cells were washed twice with HBSS and maintained in DMEM medium.

To determine the FRET efficiency of the DVL3 FlAsH III sensor, coverslips with the cells were mounted using an Attofluor holder (Molecular Probes) and placed on a Zeiss inverted microscope (Axiovert200), equipped with 63x oil objective lens and a dual-emission photometric system (Till Photonics). Cells were excited at 436 ± 10 nm using a frequency of 10 Hz with 40 ms illumination time out of a total of 100 ms. Emission of ECFP (480 ± 20 nm) and FlAsH (535 ± 15 nm), and the FRET ratio (FlAsH/ECFP) were monitored simultaneously. Fluorescence signals were detected by photodiodes and digitalized using an analogue-digital converter (Digidata 1440A, Axon Instruments). Fluorescence intensities data were acquired using Clampex software (Axon Instruments). During measurements, cells were maintained in imaging buffer; 5 mM of 2,3-dimercapto-1-propanol (also called British anti-Lewisite – BAL) was added to the cells approximately 40 s after the recording started. Recovery of ECFP fluorescence was monitored over time and FRET efficiency was calculated as described [36]. One independent experiment represents approximately 4–6 repeats (i.e. single-cell FRET signals) for each condition.

Site directed mutagenesis

The mutagenesis reactions were performed using the QuikChange II XL Site-Directed Mutagenesis Kit following the manufacturer's instructions (Agilent Technologies, #200521). All mutations described in this study were verified by Sanger sequencing. Following primers were used:

DVL3 S268A forw	ATCTCCATTGTGGACCAAGCCAACGAGCGTGGTG ACGGC
DVL3 S268A rev	GCCGTCACCACGCTCGTTGGCTTGGTCCACAATGGAGAT
DVL3 S311A for	ATCAACTTTGAGAACATGGCTAATGACGATGCAG TCCGG
DVL3 S311A rev	CCGGACTGCATCGTCATTAGCCATGTTCTCAAAGTTGAT
DVL3 S268E forw	ATCTCCATTGTGGACCAAGAGAACGAGCGTGGTG ACGGC
DVL3 S268E rev	GCCGTCACCACGCTCGTTCTCTTGGTCCACAATGGAGAT
DVL3 S311E	ATCAACTTTGAGAACATGGAGAATGACGATGCAG

Materials and methods *(Continued)*

forw	TCCGG
DVL3 S311E rev	CCGGACTGCATCGTCATTCTCCATGTTCTCAAAGTTGAT
Dvl2 S281E forw	TACAACTTCCTGGGTATCGAGATTGTTGGCCAGAGCAAT
Dvl2 S281E rev	ATTGCTCTGGCCAACAATCTCGATACCCAGGAAGTTGTA
Dvl2 S298E forw	GGCGGCATCTACATTGGCGAGATCATGAAGGGTG GGGCC
Dvl2 S298E rev	GGCCCCACCCTTCATGATCTCGCCAATGTAGATGCCGCC

Rescue assays

For transfection HEK293 T-REx (WT) and DVL1/2/3-null HEK293 T-REx cells (D1/2/3 TKO HEK293 T-REx) cells were seeded at 200000 cell/well in 24 well plate density. Next day, the cells were transfected using polyethylenimine (PEI) in concentration 1 μg/ml and pH 7.4 and the PEI ratio 6 μl of PEI/ 1 μg DNA. Mixture of transfected plasmids (10 ng of indicated DVL3 plasmid, 100 ng of Super8X TopFlash construct and 100 ng of pRLTKluc (Renilla) luciferase construct per well, the total amount of DNA was equalized by pcDNA3.1 to 400 ng DNA/well) and PEI was diluted separately in plain DMEM (DMEM without FBS, L-glutamine and antibiotics). After 6 h, medium containing transfection mix was removed and exchanged by complete DMEM medium. In all experiments cells were treated by 1 μM porcupine inhibitor LGK974 (Stem RD, 974–02) to reduce the autocrine secretion of all Wnt ligands and recombinant human R-spondin1 250 ng/ml (PeproTech, 120–38). For stimulation of the cells, Wnt3a recombinant protein was used for 14 h in concentration 80 ng/ml (R&D Systems, 5036-WN-CF). Control stimulations were done with 0.1% BSA in PBS. Samples were then analyzed by TopFlash reporter assay and WB.

Protein expression and purification

Human DVL2 PDZ domain (aa 265–361) and its phosphomimicking mutant (S281E + S298E) were cloned into pET vector with N-terminal His_6-tag, a lipoyl domain tag and a TEV cleavage site. All proteins were expressed in BL21-DE3(RIL) cells. For NMR studies, cells were grown in minimal medium (M9) supplemented by $^{15}NH_4Cl$ (1 g/l) and/or $^{13}C_6$ glucose (2 g/l) and induced with 0.5 mM IPTG for 24 h at 16 °C. Proteins were purified as described before [24] and stored in 20 mM Hepes (pH 6.8) and 50 mM KCl for NMR studies. The DVL3_C peptide 698–716 with phosphorylated S700 (DVL_C) used in NMR studies was purchased from KareBay Biochem, Inc. (New Jersey, USA).

NMR spectroscopy

NMR experiments were carried out at CEITEC Josef Dadok National NMR Centre on a 700 MHz Bruker Avance III spectrometers equipped with $^{1}H/^{13}C/^{15}N$ TCI cryogenic probe head with *z*-axis gradients. Chemical shift assignments of the phosphomimicking PDZ mutant (S281E + S298E) were obtained automatically using 4D-CHAINS technology [37] as described before for the wt PDZ domain [24]. NMR titrations were performed in series of ^{1}H-^{15}N HSQC spectra using 100 μM of ^{15}N-labeled protein (wild type or phosphomimicking mutant) and increasing amounts of DVL_C (stock concentration of 800 μM).

Mass spectrometry

In gel digestion

Immunoprecipitates were separated on SDS-PAGE gel electrophoresis, fixed with acetic acid in methanol and stained with Coomassie brilliant blue for 1 h. Corresponding 1D bands were excised. After destaining, the proteins in gel pieces were incubated with 10 mM DTT at 56 °C for 45 min. After removal of DTT excess samples were incubated with 55 mM IAA at room temperature in darkness for 30 min, then alkylation solution was removed and gel pieces were hydrated for 45 min at 4 °C in digestion solution (5 ng/μl trypsin, sequencing grade, Promega, in 25 mM AB). The trypsin digestion proceeded for 2 h at 37 °C on Thermomixer (750 rpm; Eppendorf). Subsequently, the tryptic digests were cleaved by chymotrypsin (5 ng/μl, sequencing grade, Roche, in 25 mM AB) for 2 h at 37 °C. Digested peptides were extracted from gels using 50% ACN solution with 2.5% formic acid (FA) and concentrated in speedVac concentrator (Eppendorf). The aliquot (1/10) of concentrated sample was transferred to LC-MS vial with already added polyethylene glycol (PEG; final concentration 0.001%, [38] and directly analyzed by LC-MS/MS for protein identification.

Phosphopeptide enrichment

The rest of the sample (9/10) was used for phosphopeptide analysis. Sample was diluted with acidified acetonitrile solution (80% ACN, 2% FA). Phosphopeptides were enriched using Pierce Magnetic Titanium Dioxide Phosphopeptide Enrichment Kit (Thermo Scientific, Waltham, Massachusetts, USA) according to manufacturer protocol and eluted into LC-MS vial with already added PEG (final concentration 0.001%). Eluates were concentrated under vacuum and then dissolved in water and 0.6 μl of 5% FA to get 12 μl of peptide solution before LC-MS/MS analysis.

LC-MS/MS analysis

LC-MS/MS analyses of peptide mixture were done using RSLCnano system connected to Orbitrap Elite hybrid spectrometer (Thermo Fisher Scientific) with ABIRD (Active Background Ion Reduction Device; ESI Source Solutions) and Digital PicoView 550 (New Objective) ion

source (tip rinsing by 50% acetonitrile with 0.1% formic acid) installed. Prior to LC separation, peptide samples were online concentrated and desalted using trapping column (100 μm × 30 mm) filled with 3.5 μm X-Bridge BEH 130 C18 sorbent (Waters). After washing of trapping column with 0.1% FA, the peptides were eluted (flow 300 nl/min) from the trapping column onto Acclaim Pepmap100 C18 column (3 μm particles, 75 μm × 500 mm; Thermo Fisher Scientific) by 65 min long gradient. Mobile phase A (0.1% FA in water) and mobile phase B (0.1% FA in 80% acetonitrile) were used in both cases. The gradient elution started at 1% of mobile phase B and increased from 1 to 56% during the first 50 min (30% in the 35th and 56% in 50th min), then increased linearly to 80% of mobile phase B in the next 5 min and remained at this state for the next 10 min. Equilibration of the trapping column and the column was done prior to sample injection to sample loop. The analytical column outlet was directly connected to the Digital PicoView 550 ion source.

MS data were acquired in a data-dependent strategy selecting up to top 10 precursors based on precursor abundance in the survey scan (350–2000 m/z). The resolution of the survey scan was 60,000 (400 m/z) with a target value of 1×10^6 ions, one microscan and maximum injection time of 200 ms. High resolution (15,000 at 400 m/z) HCD MS/MS spectra were acquired with a target value of 50,000. Normalized collision energy was 32% for HCD spectra. The maximum injection time for MS/MS was 500 ms. Dynamic exclusion was enabled for 45 s after one MS/MS spectra acquisition and early expiration was disabled. The isolation window for MS/MS fragmentation was set to 2 m/z.

Data analysis

The analysis of the mass spectrometric RAW data was carried out using the Proteome Discoverer software (Thermo Fisher Scientific; version 1.4) with in-house Mascot (Matrixscience; version 2.4.1) search engine utilization. MS/MS ion searches were done against in-house database containing expected protein of interest with additional sequences from cRAP database (downloaded from http://www.thegpm.org/crap/). Mass tolerance for peptides and MS/MS fragments were 7 ppm and 0.03 Da, respectively. Oxidation of methionine, deamidation (N, Q) and phosphorylation (S, T, Y) as optional modification, carbamidomethylation of C as fixed modification, TrypChymo enzyme specifity and three enzyme miss cleavages were set for all searches. The phosphoRS (version 3.1) feature was used for preliminary phosphorylation localization. Final localization of all phosphorylations (including those with ambiguous localization) was performed by manual evaluation of the fragmentation spectra of the individual phosphopeptides. Based on the presence of individual fragments in the peptide sequence, it was decided whether the localization was accurate or not.

Quantitative information was assessed and manually validated in Skyline software (Skyline daily 3.6.1.10230). Normalization of the data was performed using the set of phosphopeptide standards (added to the sample prior phosphoenrichment step; MS PhosphoMix 1, 2, 3 Light, Sigma-Aldrich) and by non-phosphorylated peptides identified in direct analyses.

All quantitative data (peptide intensities) were processed by two approaches (see Fig. 3a). In the first approach (pipeline #1 and #3) that often resulted in the formation of phosphorylated clusters all identified phosphorylated peptides were considered. This has resulted in three categories of identification: (i) one peptide or set of peptides with one clearly localized phosphorylated site, (ii) one or set of overlapping peptides covering sequence region with two or more clearly localized phosphorylated sites (phosphorylated sites separated by a comma in the Figures) and (iii) one or set of overlapping peptides covering sequence region with two or more phosphorylated sites, but some of them are not clearly localized (phosphorylated sites separated by a dash in the Figures). Sum of intensities of phosphorylated peptides (includes different peptide sequences, charges and peptides with other modifications) was calculated for each cluster. In case of direct analysis (pipeline #1), we calculated site occupancies as percentage ratio of phosphorylated peptide intensity (summed if more than one) to total intensity (summed intensities of phosphorylated peptide(s) + corresponding non-phosphorylated peptide(s)). Sites/clusters with the site occupancy > 5% at least in one experiment are shown in Fig. 4. For pipeline #3, the difference between the phosphorylated peptide summed intensity in the kinase-induced sample and the control (in log10 scale) was calculated for each replicate and the mean from all three replicates was used for the heat map (see in Additional file 2: Figure S2). In the second approach (pipeline #2), only phosphorylated peptides with the clearly localized phosphorylation site (based on manual inspection) were considered. In case of the multiphosphorylated peptides the total intensity of the peptide was assigned to each phosphorylated site. Sum of intensities of phosphorylated peptides was calculated for each phosphorylated site. The heat map (Fig. 5) was build-up in the same way as for pipeline #3 described above. For the production of phosphoplots the sum of intensities for each phosphorylated site (pipeline #2) were log-transformed (log10) and average from 3 replicates was calculated (values under the detection limit were not included in the calculation).

Numerical data & statistics

Cluster analysis

Data of absolute peptide intensities were log-transformed (log10) because of their log-normal distribution. Log-transformed data of each replicate were standardized to control (kinase subtracted from "no kinase" control) and averaged from 3 replicates. Cluster analysis was applied both for kinases and for phosphosites; Ward's minimum variance method based on the Lance-Williams recurrence and Euclidean distance were used. Circular plot accompanied by heat map was used for visualization of complex relation among phosphosites and kinases. All analyses were performed using R software.

Other analyses

One-way ANOVA and Tukey Post tests were calculated by GraphPad Prism (GraphPad Software Inc.).

Results

Identification of TTBK2 as a kinase acting upstream of DVL

Our previous work identified interactions between DVL3, CEP164, and TTBK2 kinase [34, 39]. To examine the possible regulation of DVL by TTBK2, we tested whether this poorly characterized centrosomal kinase phosphorylates DVL3 and DVL2 in HEK293 cells. As shown in Fig. 1a, TTBK2 co-expression induced a prominent electrophoretic mobility shift for both DVL3 and DVL2 in a kinase activity-dependent manner, suggesting that TTBK2 is capable to efficiently promote phosphorylation of DVL. TTBK2 also gets auto-phosphorylated as described earlier [34]. TTBK2 has been described as a kinase that primarily resides on the mother centriole where it regulates ciliogenesis [34, 40–45]. Given the previous reports on DVL centrosomal localization and the possible DVL implication in ciliogenesis [6, 7, 10, 46] we tested whether overexpressed DVL has the capacity to displace TTBK2 from the centriole. As shown in Fig. 1b (left panel), we confirmed, in line with earlier reports, the localization of TTBK2 to the mother centriole, which is however not affected by overexpression of DVL3 (Fig. 1b, right).

Mobility-shift of DVL2 and DVL3 induced by TTBK2 (Fig. 1a) can be in principle a consequence of direct phosphorylation of DVL by TTBK2 or a consequence of activation of other DVL kinases by TTBK2. In the second scenario, the well-established DVL kinase CK1ε represents the most obvious candidate target of TTBK2. To address whether TTBK2 acts directly on DVL or rather acts as CK1ε activator we treated cells with the CK1 inhibitor PF-670462 and subsequently analyzed the electrophoretic mobility shift of DVL3. As shown in Fig. 1c, PF-670462 efficiently reduced phosphorylation induced by CK1ε but not by TTBK2, hence demonstrating that TTBK2 phosphorylation of DVL does not require CK1ε activity.

The best-defined role of DVL is the positive regulation of the Wnt/β-catenin pathway. In order to address if TTBK2 modulates this DVL function we analyzed the ability of TTBK2 to promote or to inhibit DVL-induced TCF/LEF-dependent luciferase reporter (TopFlash) in HEK293 cells. Interestingly, TTBK2 did inhibit efficiently DVL1- and DVL2-induced TopFlash activation in a kinase activity-dependent manner. Because DVL3 alone induces TopFlash very poorly (Fig. 1d, right; see also [47]) we co-transfected CK1ε and DVL3 and even in this case TTBK2 co-expression reduced TopFlash activation (Fig. 1d). In summary, we demonstrate that the robust phosphorylation of DVL mediated by TTBK2 is associated with the decreased capacity of DVL to act in the Wnt/β-catenin pathway.

Design and validation of DVL kinase panel

TTBK2 represents an additional kinase to the ever-growing list of DVL kinases. Currently, multiple kinases from diverse families have been reported to phosphorylate at least one of DVL isoforms. From the fragmented published results, it is not possible to find out the unique/general or constitutive/induced phosphorylation events and patterns that are associated with various functions of DVL in connection to individual kinases. This prompted us to perform a direct comparison of DVL phosphorylation by individual kinases. We chose human HEK293 cells, a common model for the analysis of Wnt signaling, and human DVL3 as a representative DVL protein. Previously reported DVL kinases – CK1ε [11], CK2α [15], PLK1 [9], NEK2 [10, 17], PKCδ [18], RIPK4 [19] and the newly identified TTBK2 were added to the panel. We also included mitotic kinase Aurora A that has been reported to act in the same complex with DVL [25]. These S/T kinases represent very diverse members of the protein kinase family as visualized on the kinome tree (Fig. 2a). Most kinases – except for CK2α and PKCδ – could trigger electrophoretic mobility shift of DVL3 when co-expressed with DVL3 (Fig. 2b).

Pipelines for the generation of the phosphorylation map of DVL3

Following validation of our experimental system (Fig. 2) we performed a global analysis of DVL3 phosphorylation events. In three independent experiments, FLAG-DVL3 was overexpressed in HEK293 cells, with or without the studied kinase, immunoprecipitated using anti-FLAG antibody, separated on SDS-PAGE and stained with Coomassie Brilliant Blue (see Additional file 1: Figure S1 for the gels used in this study). Bands corresponding to DVL3 were cut out and after Tryp-Chymo digestion 1/10 of the sample was analyzed directly (see pipeline #1 in Fig. 3a) and the remaining 9/10 of the

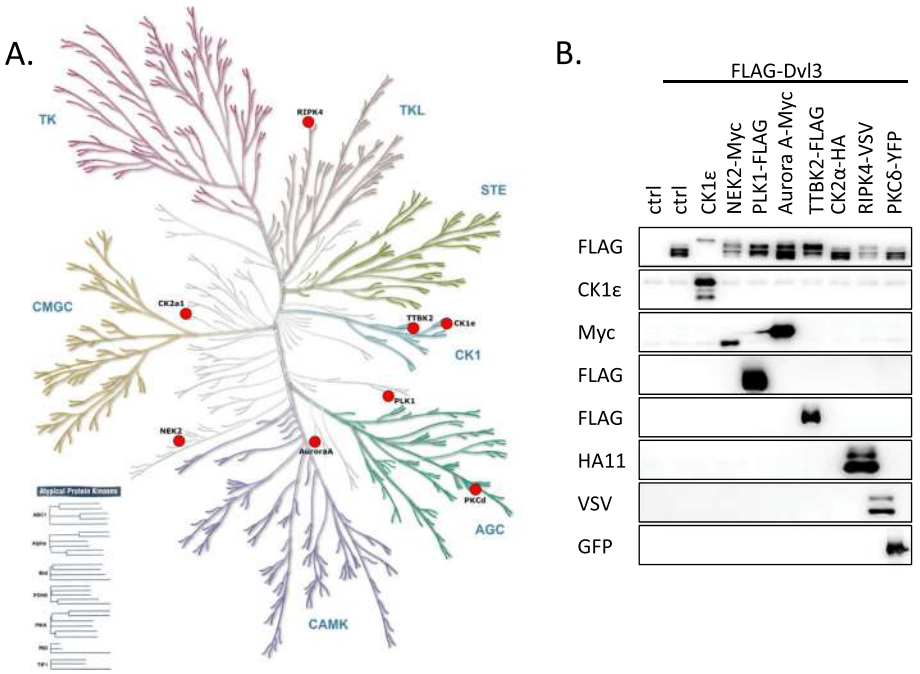


Fig. 2 Validation of the panel of DVL3 kinases. **a**: Visualization of the kinases used in this study in the phylogenetic kinome tree (http://www.kinhub.org/kinmap/). The individual kinases are representatives of distant kinase groups except for CK1ε and TTBK2 that are members of CK1 superfamily. **b**. HEK293 cells were transfected by plasmids encoding FLAG-DVL3 and the indicated kinase. Ability of individual kinases to promote DVL3 phosphorylation detected as the electrophoretic mobility shift on WB was assayed. Alpha-tubulin was used as a loading control

peptide mixture was enriched for the phosphorylated peptides by TiO_2 (see pipelines #2 and #3). All samples were subsequently analyzed by liquid chromatography coupled to mass spectrometry (LC-MS/MS).

Peptides with accurately characterized phosphorylated site(s) were identified alongside with peptides of ambiguous phospholocalization, as it is common to proteomic studies. Therefore, in the accurate position focused analysis (pipeline #2) only a subgroup of peptides with clearly identified phosphorylated positions (based on manual inspection) were included. In the overall quantitative analysis (pipeline #3) all phosphorylated peptides were processed despite the uncertainty of phosphoresidues and as such were presented as "phosphorylated clusters". For the detailed information on the quantification in these two pipelines see Materials and Methods.

Combination of the above-mentioned approaches allowed us to cover more than 95% of the DVL3 sequence across all experiments (Fig. 3b); for sequence coverage in individual samples see Additional file 6: Table S2.

Phosphorylation map of DVL3: site occupancy of phosphorylated sites

To assess the site occupancy, we utilized direct analysis of the samples without any enrichment (the experimental pipeline #1, Fig. 3a) which allowed us to detect phosphorylated and non-phosphorylated peptides corresponding to the same position(s). Subsequently we calculated the approximate occupancy of the selected phosphorylated sites, i.e. % of DVL3 molecules phosphorylated at given S or T residue(s).

We calculated site occupancies as the percentage ratio of phosphorylated peptide intensity (or sum of intensities if more than one) covering individual phosphosites or clusters to total intensity (phosphorylated peptide(s) + corresponding non-phosphorylated peptide(s)). Phosphorylated sites/clusters with > 5% site occupancy in one sample are plotted in Fig. 4. For three clusters (S232–S244, T608–S612, S622–S630) we were not able to detect matching non-phosphorylated peptides and therefore these sites were not included in the site occupancy calculations.

We observed an increase in the phosphorylated site occupancy (above 5%) after overexpression of at least one protein kinase in all but one [14] cases (Fig. 4). The notable exception was the phosphorylated cluster S192, S197, whose phosphorylation decreased in the presence of CK1ε and to a lesser extent in the presence of TTBK2, PLK1 and NEK2. Phosphorylation of three sites was induced selectively only by one kinase; S280 and S407 were induced only by NEK2, and S268 was induced only by CK1ε. In the remaining 11 sites/clusters, two or three kinases were capable to trigger

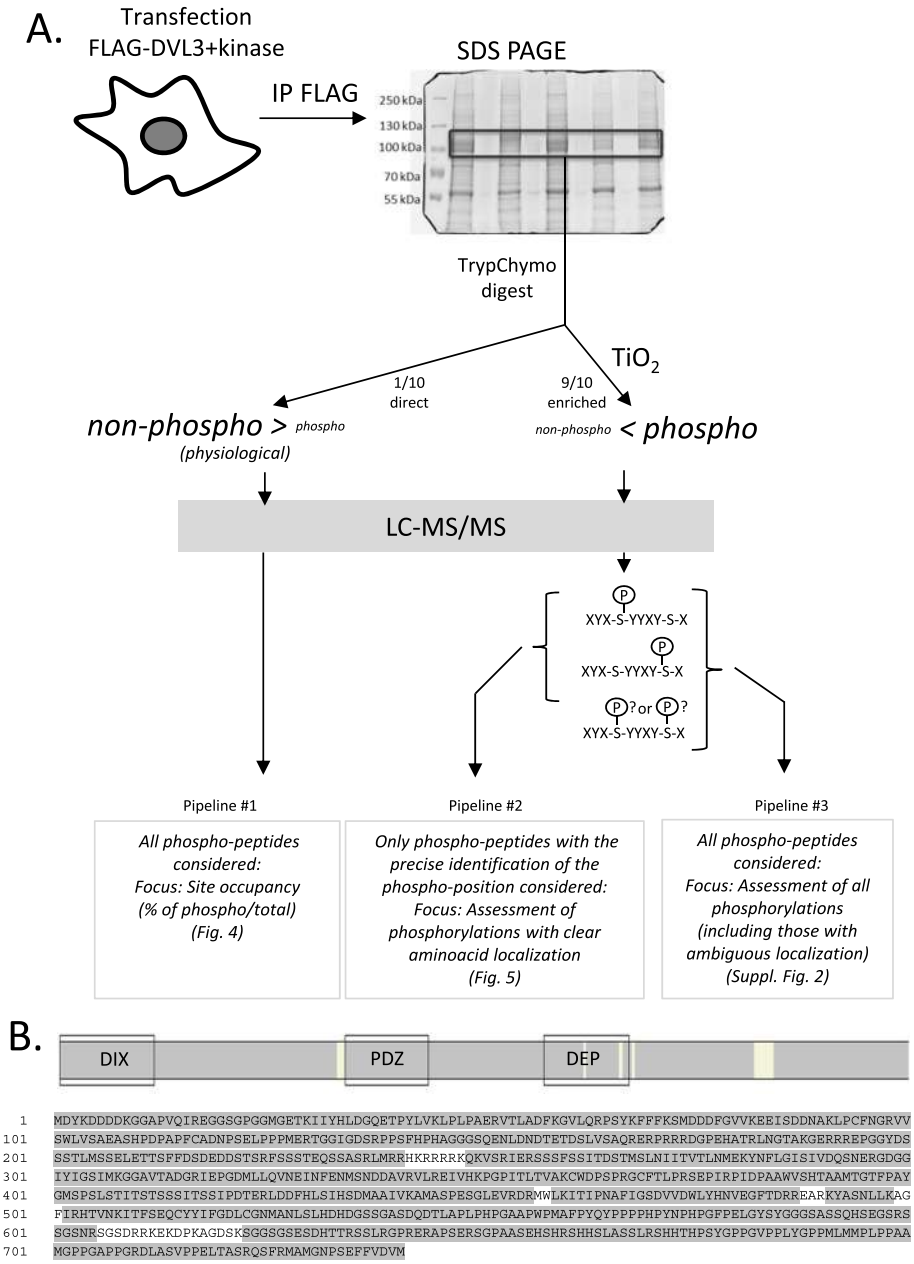


Fig. 3 Experimental design. **a**: FLAG-DVL3 was overexpressed (with or without kinase) in HEK293 cells. After cell lysis DVL3 was immunoprecipitated using anti-FLAG antibody. Immunoprecipitates were separated on SDS-PAGE gel electrophoresis, stained with Coomassie brilliant blue and the 1D bands corresponding to DVL3 were excised, digested with trypsin and subsequently cleaved by chymotrypsin. In the pipeline #1, the aliquot (1/10) of concentrated sample was directly analyzed by LC-MS/MS in order to analyze site occupancy of the abundant phosphorylated sites. The rest of the sample was enriched for phosphorylated peptides using TiO_2 and analyzed by LC-MS/MS to obtain detailed information of DVL3 phosphorylation status. Data from LC-MS/MS were searched, manually validated in Skyline software and further processed by two approaches. In the first approach (pipeline #2) only phosphorylated peptides with the clearly localized phosphorylated site (based on manual inspection of spectra) were considered. In the second approach (pipeline #3) we have considered all phosphorylated peptides that in some cases resulted in the formation of "clusters" of phosphorylated sites. **b**: The overall sequence coverage of DVL3 across all kinases and replicates. Regions of DVL3 covered by the peptides detected (Mascot score > 20) in any of the MS/MS analyses are highlighted in grey. For sequence coverage in individual samples see Additional file 6: Table S2

phosphorylation: CK1ε/TTBK2 at S350 and at the cluster S512, S513, S516; NEK2/PLK1 at S263; NEK2/Aurora A at S469 (in the single replicate); NEK2/PLK1/CK1ε at S505; NEK2/CK1ε/TTBK2 (and to some extent PLK1) at S311, and at the clusters S203–S209 and S633–S643. Three other clusters – cluster S125, T133, S135, S137, cluster S175, S176 and cluster S559, S566, S567, S570 showed relatively high occupancy in the

control but were further phosphorylated by several kinases, including NEK2, CK1ε, TTBK2 and PLK1.

Phosphorylation map of DVL3: detailed analysis after phospho-enrichment

In order to analyze phosphorylation of DVL3 in depth, we enriched the phosphorylated peptides by TiO_2. This approach, commonly utilized for detailed screening of protein phosphorylation, allows detection of less abundant phosphopeptides. Phospho-enrichment data were processed in two ways (Fig. 3a). Primarily, we assessed qualitatively and quantitatively only phosphopeptides with clearly identified (validated by manual inspection of MS/MS data) phosphorylation site(s) (the experimental pipeline #2, Fig. 3a). By this approach we detected 88 unique phosphorylation sites in DVL3. Peptide intensities (Additional file 5: Table S1) were compared with the DVL3-only control dataset in order to express the relative increase/decrease in the phosphorylation of each site. Data from all three replicates presenting the relative change to control are summarized as a heat map in Fig. 5. Most phosphorylated sites with the largest relative increase were identified after induction by NEK2, CK1ε and TTBK2.

In order to assess the impact of excluding phosphopeptides with ambiguous phosphosite localization on quantitative changes, we processed all phosphorylated peptides according to the pipeline #3 (Fig. 3a). Intensities of the phosphorylated clusters using this approach are shown as a heat map in Additional file 2: Figure S2, and illustrate that for 15 clusters analyzed, they match very well the intensities of individual phosphorylated sites presented in Fig. 5. Comparison of results in Fig. 5 and Additional file 2: Figure S2, however, also identified cases where individual phosphorylated sites within the cluster displayed a distinct behavior. Namely, in the cluster T106–S140 the NEK2-induced phosphorylation of T106, S112 and S116 was masked by high intensity constitutive phosphorylation of S125, T133 and S137. In the phosphorylated cluster S202–S209, phosphorylation was decreased for three sites – S202, S203 and S204 – in case of CK1ε and TTBK2 whereas NEK2, on the contrary, induced further phosphorylation of S204. Similarly, for the cluster S598–S612 induced by NEK2, CK1ε, TTBK2 and PKCδ we mapped the activity of NEK2 predominantly to S601, S603 and S605 whereas CK1ε phosphorylated mainly to S611 and S612.

The quantitative data presented in Fig. 5 (accurate positions) and Additional file 2: Figure S2 (clusters) have been combined and analyzed in order to correlate individual kinases with individual phosphorylated sites/clusters. Unbiased cluster analysis (Additional file 3: Figure S3) groups CK1ε and TTBK2, both from the CK1 family, with NEK2, whereas all the remaining kinases form a second group. There are multiple residues that are phosphorylated only by NEK2, which further distinguishes it from CK1ε and TTBK2. Interestingly, CK1ε and TTBK2 behave very similarly and they can be best resolved by the phosphorylation at position S268 that is CK1ε-specific. This is in very good agreement with the site occupancy analysis shown in Fig. 4.

Phosphorylation map of DVL3: analysis of the phosphorylated clusters

The phosphorylated clusters visualized in Additional file 2: Figure S2 can, in principle, represent mixtures of peptides phosphorylated at distinct sites or true multiphosphorylated signatures with possible biological function. To get a better insight, we analyzed in detail the peptides phosphorylated at 3 or more sites. We found 9 peptide families fulfilling these requirements (Fig. 6). Out of these only one cluster – S112, S116, S125, T133, S135, S137 – was not induced by any of the kinases. The remaining 8 clusters were induced by one or several kinases – 3 only by NEK2, 2 mainly by CK1ε and TTBK2 and 3 by NEK2 and one or more additional kinases.

These phosphorylated clusters were detected only in the intrinsically disordered regions of DVL3. The phosphorylated motifs are well conserved and found also in DVL1 and DVL2 except for S622–630 and S636–S643 (Fig. 6, right). This may suggest that the function of multisite phosphorylation of these motifs in the regulation of DVL is also conserved.

Comparison of individual pipelines and validation by phospho-specific antibodies

In our study, we used several sample and data processing pipelines (see Fig. 3). As the last step we compared the individual pipelines (i) between each other and (ii) with several phosphorylation-specific antibodies raised against phosphorylated DVL3 peptides. We probed the samples described in Fig. 2b with antibodies against phosphorylated-S280-DVL3 (pS280) [10], pS643 [20], pS697 [10] and the newly generated anti-pS192. As shown in Fig. 7 a-d that combines Western blots and MS/MS data from pipelines #1 and #2, the signal of phospho-specific antibodies partially matches the changes observed by mass spectrometry. The signal of pS192-DVL3 decreased after CK1ε, PLK1 and TTBK2 co-expression whereas pS280, pS643 and pS697 increased mostly by NEK2. However, some discrepancies were observed, especially for low intensity signals, namely, all three previously validated antibodies pS280, pS643 and pS697 do detect higher phosphorylation after Aurora A co-expression that was not detected by LC-MS/MS. On the other hand, increase in S643 phosphorylation probed for TTBK2 by all MS/MS pipelines was not detected by anti-pS643-DVL3 antibody.

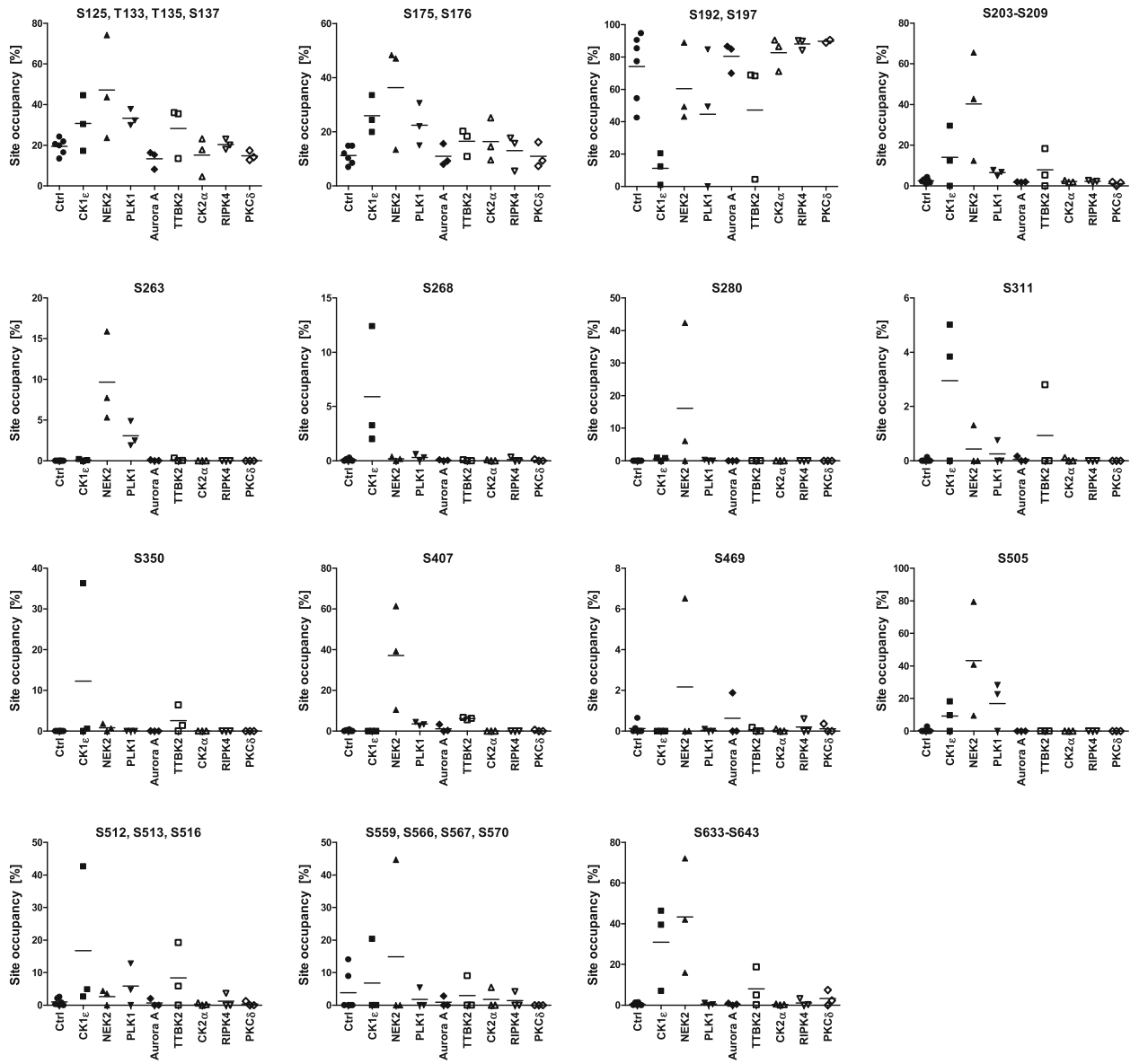
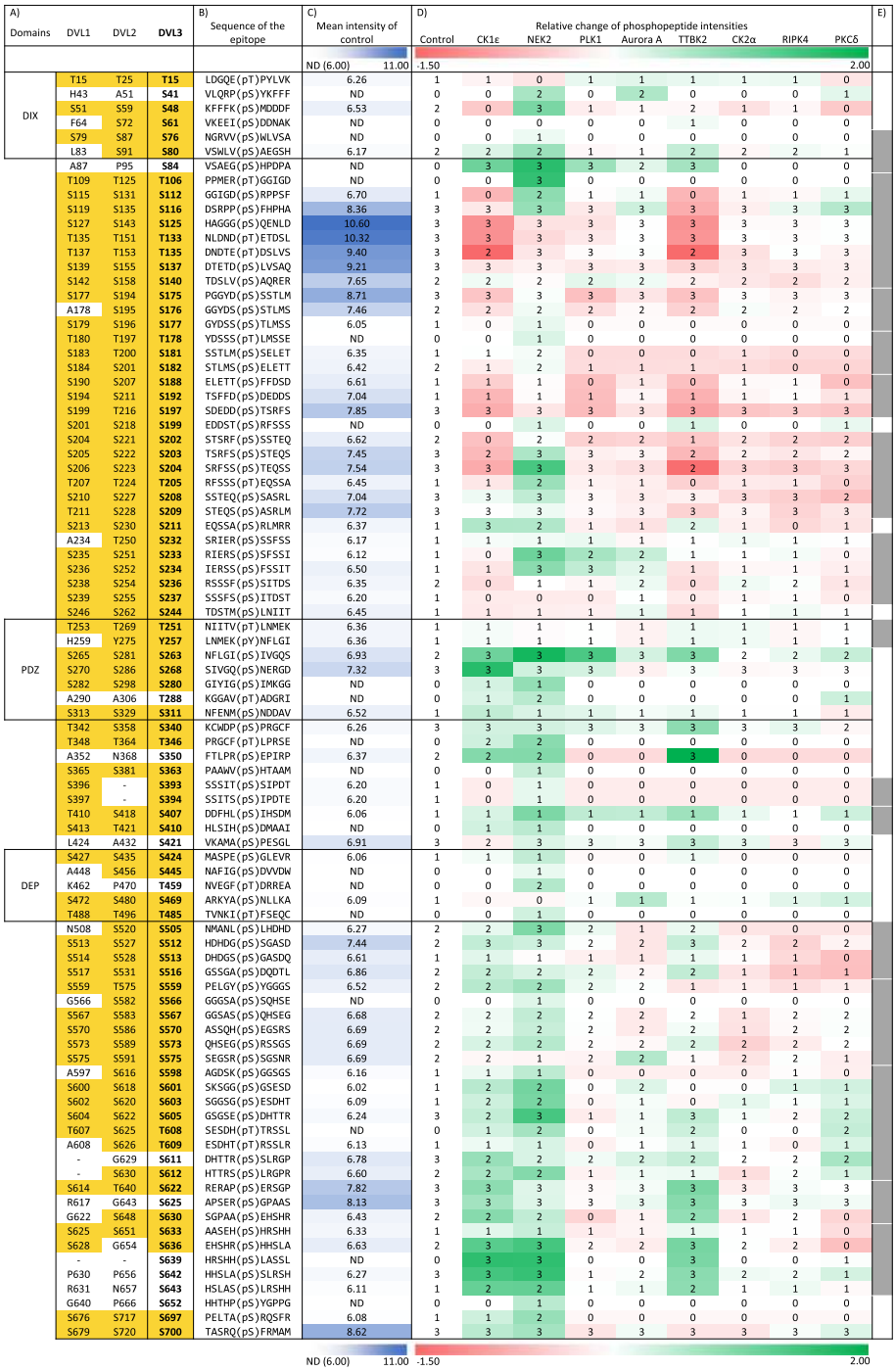


Fig. 4 Site occupancy of the abundant phosphorylated sites. Site occupancy analysis was performed according to the pipeline #1 in Fig. 3a. Fifteen phosphorylated peptides or clusters that were phosphorylated in more than 5% at least in one replicate are plotted. Graphs present individual data points from three biological replicates (two controls/biological sample) and the mean values (horizontal line)

Phosphoplot – a tool for visualization of complex phosphorylation patterns

The heat map data visualization highlights phosphorylation differences in comparison to control but does not provide a full and intuitive picture combining absolute intensities, their changes and position of individual phosphorylated peptides. To address this, we designed a visualization diagram, that we refer to as "phosphoplot". In the phosphoplot all S/T from the primary sequence of the analyzed protein are shown. The intensities of phosphorylated peptides in the control and experimental conditions are plotted. As such, the phosphoplot combines information on absolute peptide intensities, experimental differences and positional information, including non-phosphorylated sites. The phosphoplots of DVL3 for individual kinases based on data from pipeline #2 are shown in Fig. 8. Phosphoplot inspection identified regions of DVL3 without detected phosphorylation despite being S/T rich – such as T365–T392 between PDZ and DEP domains, and on the contrary, regions between DIX and PDZ domains that are highly constitutively phosphorylated. It also clearly identifies the uniquely phosphorylated sites for each kinase.



A) Domains	DVL1	DVL2	DVL3	B) Sequence of the epitope	C) Mean intensity of control	D) Relative change of phosphopeptide intensities: Control	CK1ε	NEK2	PLK1	Aurora A	TTBK2	CK2α	RIPK4	PKCδ	E)
					ND (6.00) 11.00	-1.50								2.00	
DIX	T15	T25	**T15**	LDGQE(pT)PYLVK	6.26	1	1	0	1	1	1	1	1	0	
	H43	A51	**S41**	VLQRP(pS)YKFFF	ND	0	0	2	0	2	0	0	0	1	
	S51	S59	**S48**	KFFFK(pS)MDDDF	6.53	2	0	3	1	1	2	1	1	0	
	F64	S72	**S61**	VKEEI(pS)DDNAK	ND	0	0	0	0	0	1	0	0	0	
	S79	S87	**S76**	NGRVV(pS)WLVSA	ND	0	0	1	0	0	0	0	0	0	
	L83	S91	**S80**	VSWLV(pS)AEGSH	6.17	2	2	2	1	1	2	2	2	1	
	A87	P95	**S84**	VSAEG(pS)HPDPA	ND	0	3	3	3	2	3	0	0	0	
	T109	T125	**T106**	PPMER(pT)GGIGD	ND	0	0	3	0	0	0	0	0	0	
	S115	S131	**S112**	GGIGD(pS)RPPSF	6.70	1	0	2	1	1	0	1	1	1	
	S119	S135	**S116**	DSRPP(pS)FHPHA	8.36	3	3	3	3	3	3	3	3	3	
	S127	S143	**S125**	HAGGG(pS)QENLD	10.60	3	3	3	3	3	3	3	3	3	
	T135	T151	**T133**	NLDND(pT)ETDSL	10.32	3	3	3	3	3	3	3	3	3	
	T137	T153	**T135**	DNDTE(pT)DSLVS	9.40	3	2	3	3	3	2	3	3	3	
	S139	S155	**S137**	DTETD(pS)LVSAQ	9.21	3	3	3	3	3	3	3	3	3	
	S142	S158	**S140**	TDSLV(pS)AQRER	7.65	2	2	2	2	2	2	2	2	2	
	S177	S194	**S175**	PGGYD(pS)SSTLM	8.71	3	3	3	3	3	3	3	3	3	
	A178	S195	**S176**	GGYDS(pS)STLMS	7.46	2	2	2	2	2	2	2	2	2	
	S179	S196	**S177**	GYDSS(pS)TLMSS	6.05	1	0	1	0	0	0	0	0	0	
	T180	T197	**T178**	YDSSS(pT)LMSSE	ND	0	0	1	0	0	0	0	0	0	
	S183	T200	**S181**	SSTLM(pS)SELET	6.35	1	1	2	0	0	0	1	0	0	
	S184	S201	**S182**	STLMS(pS)ELETT	6.42	2	1	2	1	1	1	1	0	0	
	S190	S207	**S188**	ELETT(pS)FFDSD	6.61	1	1	1	0	1	0	1	1	0	
	S194	S211	**S192**	TSFFD(pS)DEDDS	7.04	1	1	1	1	1	1	1	1	1	
	S199	T216	**S197**	SDEDD(pS)TSRFS	7.85	3	3	3	3	3	3	3	3	3	
	S201	S218	**S199**	EDDST(pS)RFSSS	ND	0	0	1	0	0	1	0	0	1	
	S204	S221	**S202**	STSRF(pS)SSTEQ	6.62	2	0	2	2	2	1	2	2	2	
	S205	S222	**S203**	TSRFS(pS)STEQS	7.45	3	2	3	3	3	2	2	2	2	
	S206	S223	**S204**	SRFSS(pS)TEQSS	7.54	3	3	3	3	3	2	3	3	3	
	T207	T224	**T205**	RFSSS(pT)EQSSA	6.45	1	1	2	1	1	0	1	1	0	
	S210	S227	**S208**	SSTEQ(pS)SASRL	7.04	3	3	3	3	3	3	3	3	2	
	T211	S228	**S209**	STEQS(pS)ASRLM	7.72	3	3	3	3	3	3	3	3	3	
	S213	S230	**S211**	EQSSA(pS)RLMRR	6.37	1	3	2	1	1	2	1	0	1	
	A234	T250	**S232**	SRIER(pS)SSFSS	6.17	1	1	1	1	1	1	1	1	1	
	S235	S251	**S233**	RIERS(pS)FSSSI	6.12	1	0	3	2	2	1	1	1	0	
	S236	S252	**S234**	IERSS(pS)FSSIT	6.50	1	1	3	3	2	1	1	1	1	
	S238	S254	**S236**	RSSSF(pS)SITDS	6.35	2	0	1	1	2	0	2	2	1	
	S239	S255	**S237**	SSSFS(pS)ITDST	6.20	1	0	0	0	1	0	1	1	0	
	S246	S262	**S244**	TDSTM(pS)LNIIT	6.45	1	1	1	1	1	1	1	1	1	
PDZ	T253	T269	**T251**	NIITV(pT)LNMEK	6.36	1	1	1	1	1	1	1	1	1	
	H259	Y275	**Y257**	LNMEK(pY)NFLGI	6.36	1	1	1	1	1	1	1	1	1	
	S265	S281	**S263**	NFLGI(pS)IVGQS	6.93	2	3	3	3	3	3	2	2	2	
	S270	S286	**S268**	SIVGQ(pS)NERGD	7.32	3	3	3	3	3	3	3	3	3	
	S282	S298	**S280**	GIYIG(pS)IMKGG	ND	0	1	1	0	0	0	0	0	0	
	A290	A306	**T288**	KGGAV(pT)ADGRI	ND	0	1	2	0	0	0	0	0	1	
	S313	S329	**S311**	NFENM(pS)NDDAV	6.52	1	1	1	1	1	1	1	1	1	
	T342	S358	**S340**	KCWDP(pS)PRGCF	6.26	3	3	3	3	3	3	3	3	2	
	T348	T364	**T346**	PRGCF(pT)LPRSE	ND	0	2	2	0	0	0	0	0	0	
	A352	N368	**S350**	FTLPR(pS)EPIRP	6.37	2	2	2	0	0	3	0	0	0	
	S365	S381	**S363**	PAAWV(pS)HTAAM	ND	0	0	1	0	0	0	0	0	0	
	S396	-	**S393**	SSSIT(pS)SIPDT	6.20	1	0	1	0	0	0	0	0	0	
	S397	-	**S394**	SSITS(pS)IPDTE	6.20	1	0	1	0	0	0	0	0	0	
	T410	S418	**S407**	DDFHL(pS)IHSDM	6.06	1	1	1	1	1	1	1	1	1	
	S413	T421	**S410**	HLSIH(pS)DMAAI	ND	0	1	1	0	0	0	0	0	0	
	L424	A432	**S421**	VKAMA(pS)PESGL	6.91	3	2	3	3	3	3	3	3	3	
DEP	S427	S435	**S424**	MASPE(pS)GLEVR	6.06	1	1	1	0	0	1	0	0	0	
	A448	S456	**S445**	NAFIG(pS)DVVDW	ND	0	0	1	0	0	0	0	0	0	
	K462	P470	**T459**	NVEGF(pT)DRREA	ND	0	0	2	0	0	0	0	0	0	
	S472	S480	**S469**	ARKYA(pS)NLLKA	6.09	1	0	0	1	1	1	1	1	1	
	T488	T496	**T485**	TVNKI(pT)FSEQC	ND	0	0	1	0	0	0	0	0	0	
	N508	S520	**S505**	NMANL(pS)LHDHD	6.27	2	2	3	2	1	2	0	0	0	
	S513	S527	**S512**	HDHDG(pS)SGASD	7.44	2	3	3	2	2	3	2	2	2	
	S514	S528	**S513**	DHDGS(pS)GASDQ	6.61	1	1	1	1	1	1	1	1	0	
	S517	S531	**S516**	GSSGA(pS)DQDTL	6.86	2	2	2	2	2	2	1	1	1	
	S559	T575	**S559**	PELGY(pS)YGGGS	6.52	2	2	2	2	2	1	1	1	1	
	G566	S582	**S566**	GGGSA(pS)SQHSE	ND	0	0	1	0	0	0	0	0	0	
	S567	S583	**S567**	GGSAS(pS)QHSEG	6.68	2	2	2	2	2	2	1	2	2	
	S570	S586	**S570**	ASSQH(pS)EGSRS	6.69	2	2	2	2	2	2	1	2	2	
	S573	S589	**S573**	QHSEG(pS)RSSGS	6.69	2	2	2	2	2	2	2	2	2	
	S575	S591	**S575**	SEGSR(pS)SGSNR	6.69	2	2	1	2	2	1	2	2	1	
	A597	S616	**S598**	AGDSK(pS)GGSGS	6.16	1	1	1	0	0	0	0	1	0	
	S600	S618	**S601**	SKSGG(pS)GSESD	6.02	1	2	2	0	2	0	0	1	1	
	S602	S620	**S603**	SGGSG(pS)ESDHT	6.09	1	2	2	0	1	0	1	1	1	
	S604	S622	**S605**	GSGSE(pS)DHTTR	6.24	3	2	3	1	1	3	1	2	2	
	T607	S625	**T608**	SESDH(pT)TRSSL	ND	0	1	2	0	1	2	0	0	2	
	A608	S626	**T609**	ESDHT(pT)RSSLR	6.13	1	1	1	0	1	1	1	0	1	
	-	G629	**S611**	DHTTR(pS)LRGP	6.78	3	2	2	2	2	2	2	2	2	
	-	S630	**S612**	HTTRS(pS)LRGPR	6.60	2	2	2	1	1	1	1	2	1	
	S614	T640	**S622**	RERAP(pS)ERSGP	7.82	3	3	3	3	3	3	3	3	3	
	R617	G643	**S625**	APSER(pS)GPAAS	8.13	3	3	3	3	3	3	3	3	3	
	G622	S648	**S630**	SGPAA(pS)EHSHR	6.43	2	2	2	0	1	2	1	2	0	
	S625	S651	**S633**	AASEH(pS)HRSHH	6.33	1	1	1	1	1	1	1	1	0	
	S628	G654	**S636**	EHSHR(pS)HHSLA	6.63	2	3	3	2	2	3	2	2	0	
	-	-	**S639**	HRSHH(pS)LASSL	ND	0	3	3	0	0	3	0	0	1	
	P630	P656	**S642**	HHSLA(pS)SLRSH	6.27	3	3	3	1	2	3	2	2	1	
	R631	N657	**S643**	HSLAS(pS)LRSHH	6.11	1	2	2	1	1	2	1	1	1	
	G640	P666	**S652**	HHTHP(pS)YGPPG	ND	0	0	1	0	0	0	0	0	0	
	S676	S717	**S697**	PELTA(pS)RQSFR	6.08	1	1	2	0	0	0	0	0	0	
	S679	S720	**S700**	TASRQ(pS)FRMAM	8.62	3	3	3	3	3	3	3	3	3	
					ND (6.00) 11.00	-1.50								2.00	

Fig. 5 Phosphorylation map of DVL3. All identified phosphorylation sites obtained from the pipeline #2 are visualized as a heatmap. Color intensities reflect relative change in the site phosphorylation (red – decrease, green – increase). Following additional information is also provided: **a**) Corresponding sites in human DVL1 and DVL2. Positions conserved in DVL1 and/or DVL2 either as Ser or Thr are highlighted in yellow. Position of the structured domains (DIX, PDZ and DEP) is indicated. **b**) The sequence of the phosphorylated epitope. Five amino acids before and after the identified phosphorylated site are shown. **c**) Mean absolute intensities of the phosphorylation sites in the control (DVL3 without exogenous kinase; $N = 6$) are expressed in the shades of blue. Numbers indicate decadic logarithm of the mean. ND indicated in white corresponds means "not detected". All signals lower than 1×10^6, corresponding to log value 6.0 were considered as not detected. **d**) Nine columns represent heat map of relative change of phosphorylated peptide intensities (in log10 scale) obtained for individual kinases (relative to control). Numbers in the heatmap fields (0, 1, 2, 3) indicate the number of experimental replicates with the positive identification of the given phosphorylated site. **e**) Grey boxes indicate the position of clusters of sites analyzed also according to the pipeline #3 in the Additional file 2: Figure S2

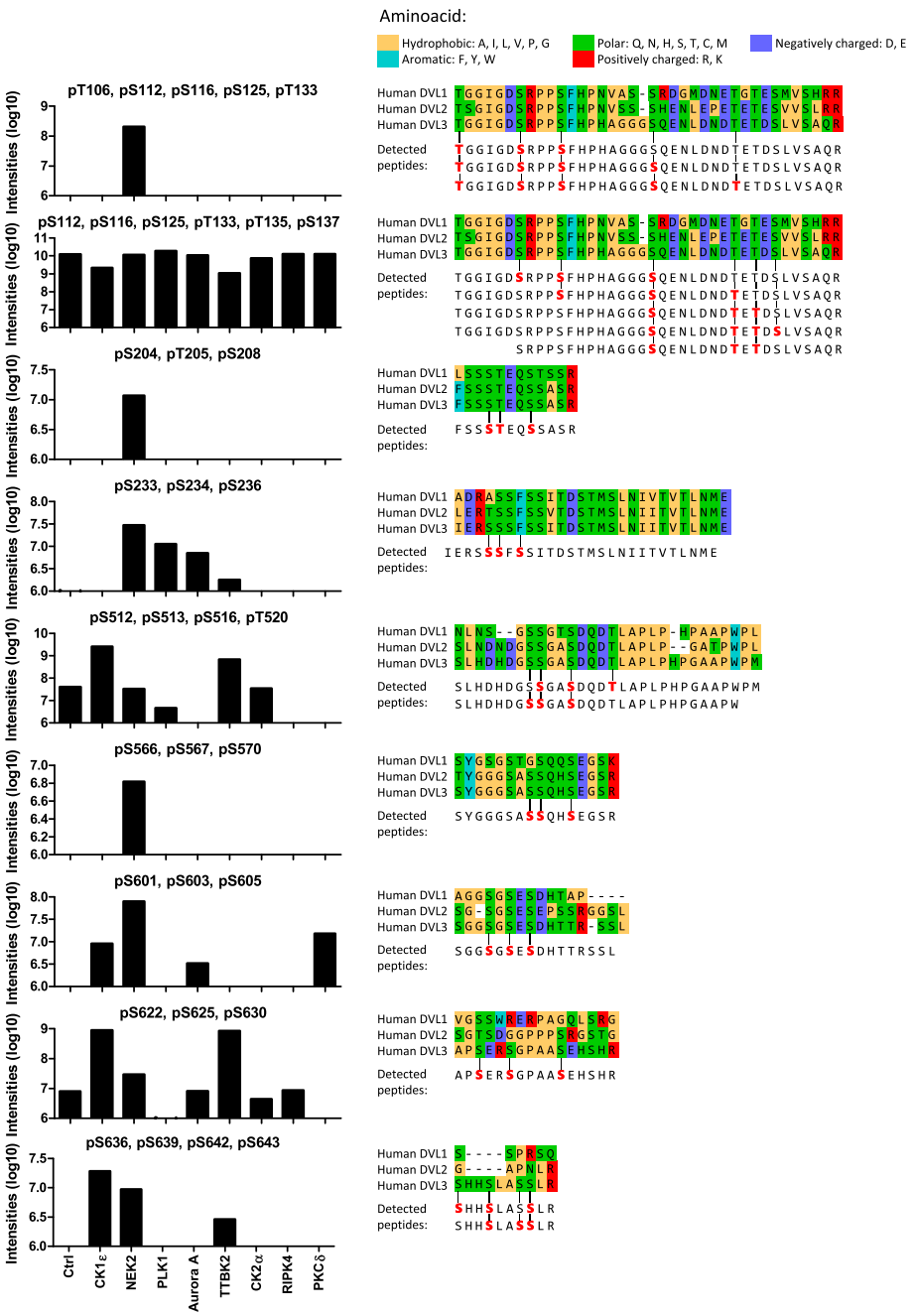


Fig. 6 Clusters phosphorylated in more than three sites. The clusters of Ser/Thr where 3 or more phosphorylated sites in one peptide were analyzed (i.e. multiphosphorylated peptides). Graphs indicate total intensities of the multiphosphorylated peptides from all three replicates; signal intensities from pipeline #1 and #3 are merged. Multiple sequence alignment shows the evolutionary conservation of the motif among individual DVL isoforms. All combinations of individual multiphosphorylated peptides detected in the cluster are indicated

Phosphorylation of S268 and S311 contributes to the activation of the Wnt/β-catenin pathway

To associate any phosphorylation patterns with the function of DVL we compared individual kinases in their capacity to induce Wnt/β-catenin-dependent transcription analyzed by the TopFlash reporter assay. In the absence of exogenous DVL3 only CK1ε and RIPK4 significantly activated TCF/LEF-driven transcription (Fig. 9a). When DVL3 was co-expressed, only CK1ε, and to a lesser extent PLK1 (non-significant trend was observed also for CK2α), could synergize with DVL3 to promote reporter activation (Fig. 9b). We could largely reproduce the reported effects: the capacity of CK1ε and RIPK4 to activate TopFlash reporter [19, 48], and the synergistic behavior of DVL3 and

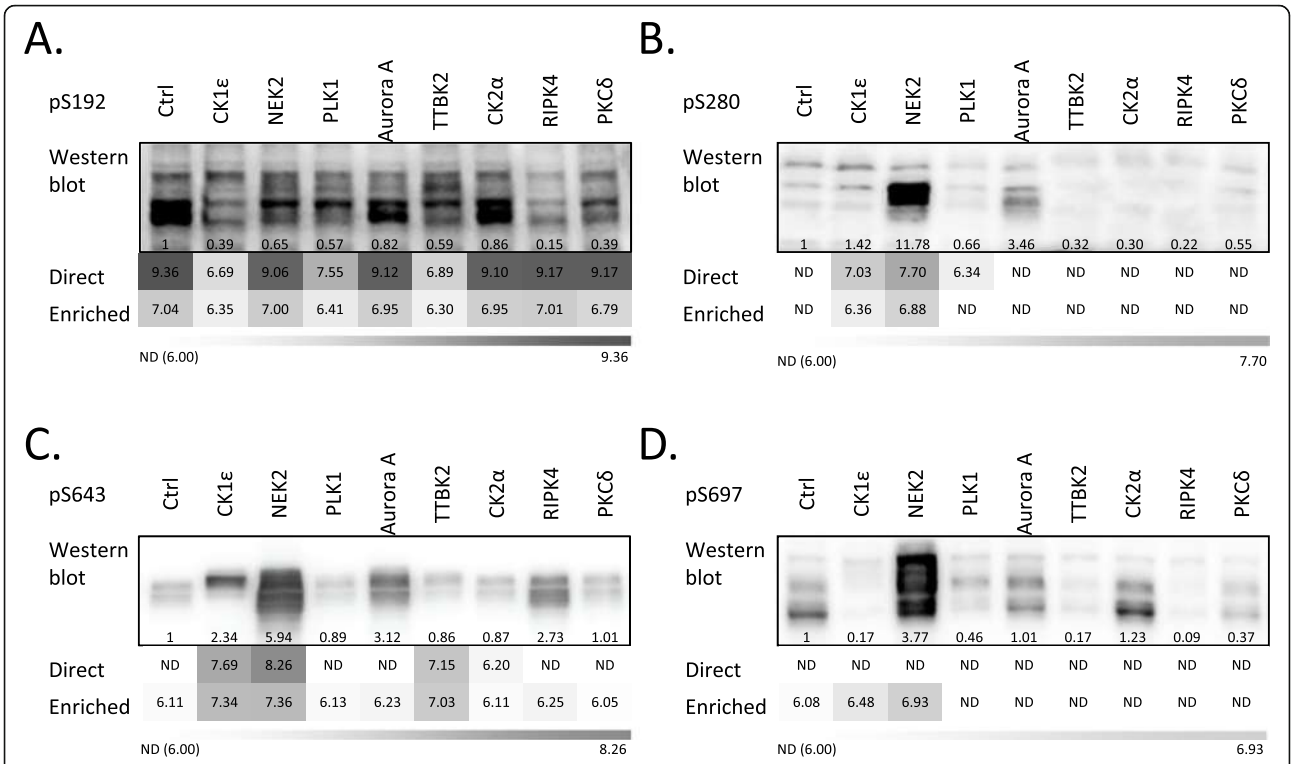


Fig. 7 Comparison of individual methods and their validation by phospho-specific antibodies. **a**-**d**. Comparison of methods. Existing phosphorylation-specific antibodies were used to visualize the level of phosphorylation of DVL3 with our kinase panel. HEK293 cells were transfected by the indicated combination of plasmids and analyzed by Western Blotting. Reactivity of individual phosphoantibodies against DVL3 phosphorylated by individual kinases is shown. Western blots are quantified using ImageJ software as absolute values for peak areas of corresponding bands and the intensities were normalized to the control. Mean intensities of the phosphorylated peptides obtained by MS/MS via pipelines #1 and #2 are shown in the shades of grey. Numbers indicate decadic logarithm of the mean peptide intensity. "Not detected" (ND) indicated in white corresponds to the signals below 1×106, i.e. 6.0. **a**. anti-pS192 (this study), **b**. anti-pS280, **c**. anti-pS643, **d**. anti-pS697

CK1ε in the TopFlash reporter [16]. Although, it has been proposed that RIPK4 activates Wnt/β-catenin also via phosphorylation of DVL [19] we have not observed any synergy with DVL3 which suggests that RIPK4 may act via other proteins in the Wnt/β-catenin pathway.

CK1ε stands unique among other kinases in its ability to efficiently induce Wnt/β-catenin downstream signaling in synergy with DVL3. We have thus analyzed the data reported in Figs. 4, 5 and Additional file 3: Figure S3 and identified S268 and S311 as candidate phosphorylation sites associated with the expression of CK1ε and activation of TopFlash reporter assay. In order to test to what extent phosphorylation of S268 and S311 participates in the activation of Wnt/β-catenin downstream signaling we mutated S268 alone or in combination with S311 to alanine and glutamic acid (S268A, S268E, S268A/S311A, S268E/S311E) and tested these DVL3 mutants for their capacity to rescue Wnt-3a signaling in DVL1/DVL2/DVL3-triple null HEK293 T-REx cells (D1/2/3 TKO cells) [47]. As shown in Fig. 9 C/D, DVL3 S268A and especially S268A/S311A are significantly less efficient in their ability to rescue Wnt-3a-induced activity. This data suggests that phosphorylation of DVL3 at S268 and S311 by CK1ε is one of the essential steps in the Wnt-3a-induced activation of β-catenin signaling.

Phosphorylation of S630-S643 mediates even localization of DVL3

Subcellular localization of DVL3 is dynamically regulated. Overexpressed DVL3, similarly to other DVL proteins, is localized in the "DVL punctae"; protein assemblies kept together via polymerization of DVL DIX domains [49]. These assemblies are dynamic and can be more compact (visible as puncta) or dissolved, resulting in the "even" distribution of DVL3 (examples of these two DVL states are presented in Fig. 10a). As shown in Fig. 10b, bottom panel, only three kinases – CK1ε, NEK2 and TTBK2 – were capable to promote even localization of DVL3.

In order to identify phosphorylation patterns associated with the even localization of DVL3 we analyzed common phosphorylation events induced by CK1ε, TTBK2 and NEK2. A cluster of the regularly spaced serines S630-S633-S636-S639-S642-S643 was the most obvious candidate for such function (see Figs. 4, 5

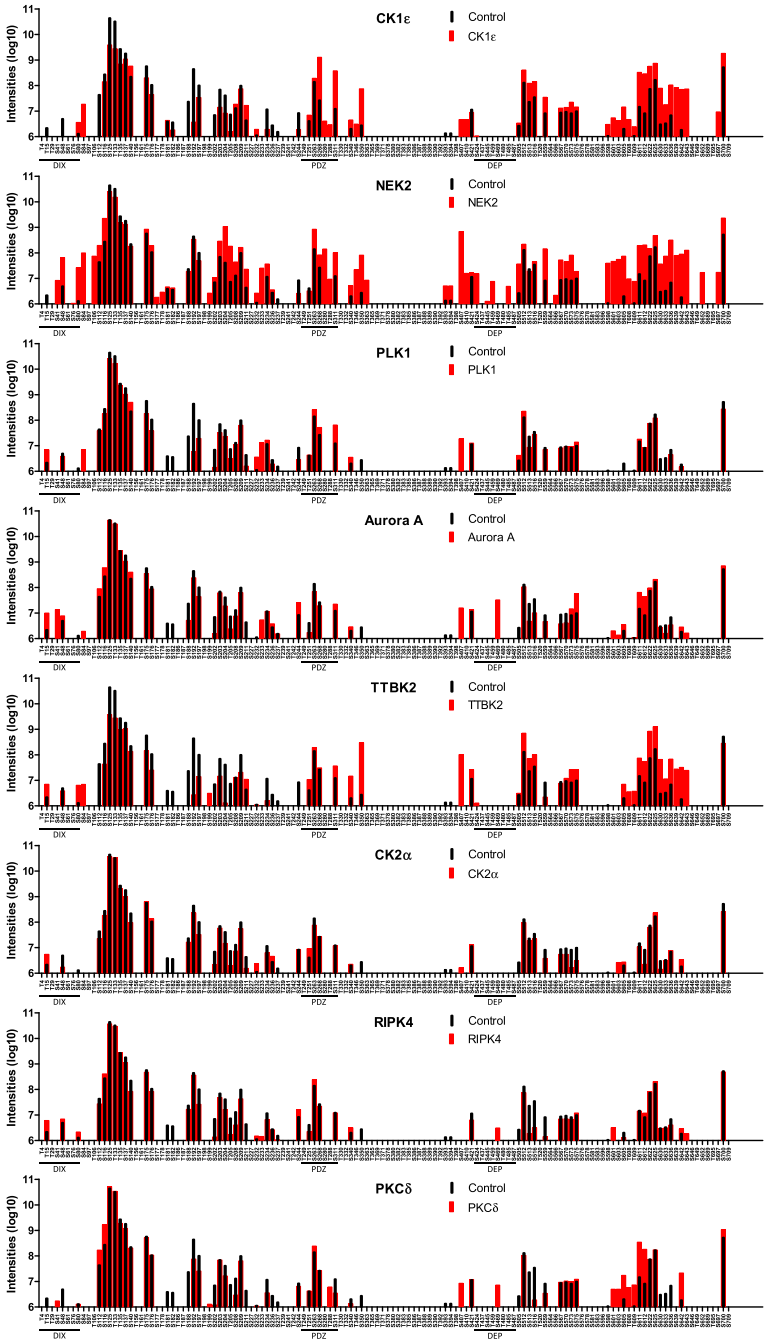


Fig. 8 Phosphoplots - phosphorylation barcodes of DVL3 with the individual kinases. Visualization of absolute intensity of phosphorylated peptides corresponding to the individual phosphorylated sites plotted on the primary sequence of DVL3 (only Ser and Thr are shown). Black bars represent a control condition, red bars the intensities in the presence of the kinase. Intensities are plotted on a log10 scale

and 6). Mutation of these residues to alanine (A) prevented the ability of CK1ε (Fig. 10c) to induce the even distribution of DVL3, in line with an earlier report [23]. Similar and even stronger effect was observed for TTBK2 suggesting that TTBK2 acts via a similar mechanism as CK1ε and that phosphorylation of S630-S643 is essential for TTBK2-induced even localization of DVL3. Surprisingly, NEK2 efficiently promoted even localization of DVL3 in all DVL3 mutants (Fig. 10c, right). This suggested that although NEK2 strongly phosphorylates S630-S643 cluster, it is able to overcome the requirement for S630-S643 phosphorylation and efficiently induce even localization of DVL3 also via a different mechanism.

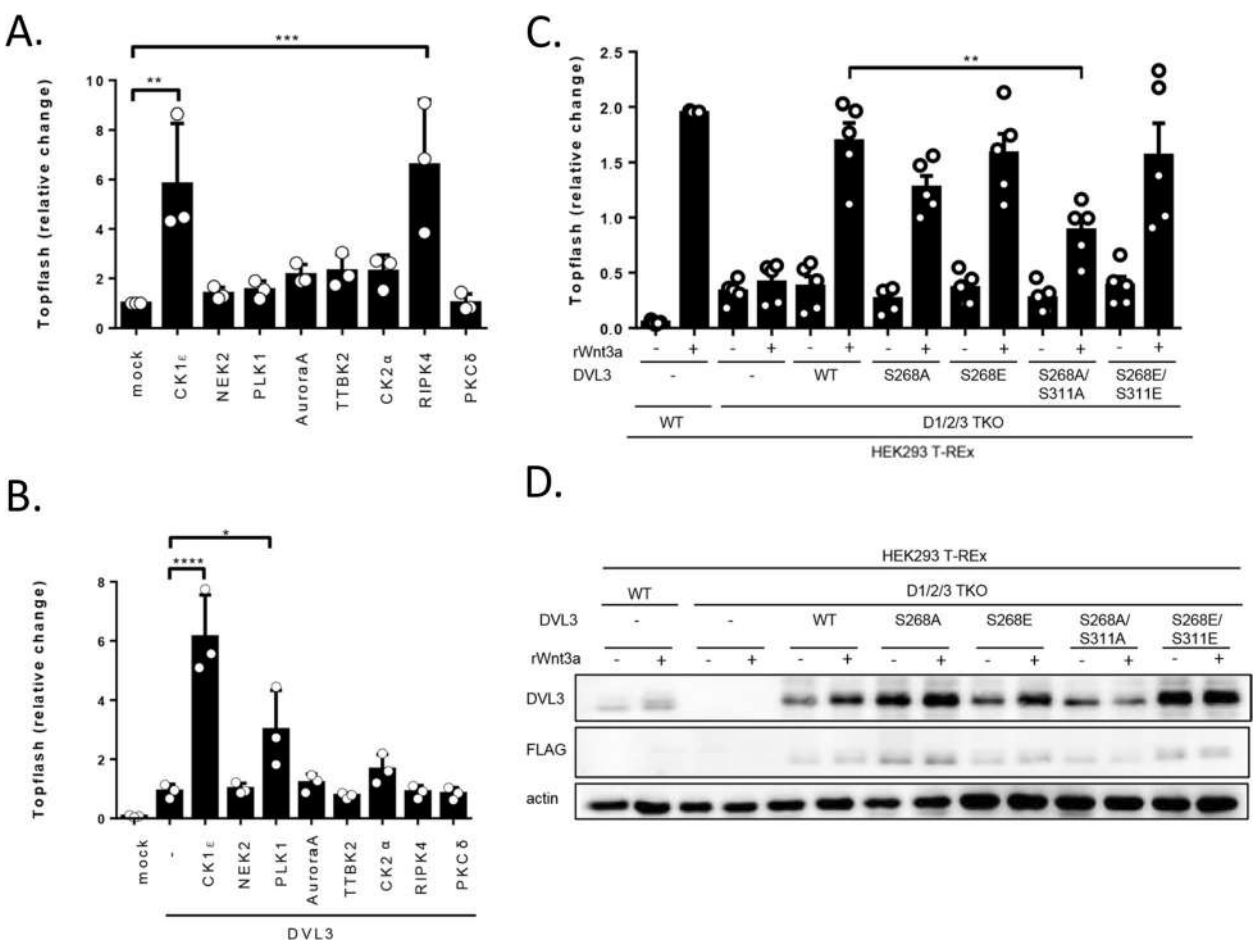


Fig. 9 Phosphorylation of S268 and S311 contributes to the activation of the Wnt/b-catenin pathway. **a**, **b**: HEK293 cells were transfected by indicated plasmids together with TopFlash and Renilla reporter plasmids. The ability of individual kinases to induce activation of Wnt/β-catenin pathway either alone (**a**) or in combination with DVL3 (**b**) was analyzed. Mean, SD and individual data points are indicated. Statistical differences were tested by One-way ANOVA and Tukey's post test (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$). **c**, **d**: Rescue experiments with DVL3 S268/S311 mutants. WT and D1/2/3 TKO HEK293 T-Rex cells were transfected as indicated and treated according to the scheme with 80 ng/ml recombinant human Wnt3a (rWnt3a). All samples were treated with 0.1 μM LGK974 inhibitor and 250 ng/ml R-SPONDIN1 and analyzed by TopFlash assay (**c**). Mean, SD and individual data points are indicated. Statistical differences were tested by paired t-test (* $p < 0.05$, ** $p < 0.01$). Samples were also used for WB analysis (**d**)

Phosphorylation of S263 and S280 by NEK2 induces open conformation of DVL3

We have recently shown a correlation between the subcellular localization of DVL3 and DVL3 conformation [24]. DVL3 forms a closed conformation via an intramolecular interaction between the C-terminus and the PDZ domain [50]. To test whether a similar mechanism can overcome the requirement for phosphorylation of S630-S643 cluster by NEK2 we identified the phosphorylation sites in the PDZ domain uniquely induced by NEK2. S263 and S280 fit the criteria as the candidate sites for functional validation (see Fig. 4 and Fig. 5). In order to address the possible role of DVL3 phosphorylation-driven regulation involving S263 and S280 in vitro, we performed functional characterization by NMR. For this purpose, we exploited the well-characterized PDZ domain of human DVL2 (aa 265–361) [51], in which S281 and S298 correspond to S263 and S280 of human DVL3. As seen in the crystal structure of the complex between DVL2 PDZ and an internal peptide ligand [52] the fully conserved S281 and S298 (Fig. 11a) reside at adjacent strands of the PDZ fold in close proximity to the peptide binding groove (Fig. 11b). From the NMR fingerprint spectra, we concluded that the phosphorylation-mimicking substitutions (S281E/S298E) did not affect the integrity of the PDZ fold (Additional file 4: Figure S4). Next, we assessed PDZ binding to a C-terminal peptide ligand derived from DVL3 (aa 698–716) by NMR titrations [24]. The residues in the binding loop region (L278 and G279) that serve as the H-bonding donors to the carboxyl sidechain of the peptide Asp sidechain undergo severe line broadening in wildtype PDZ, whereas in the phosphomimicking mutant

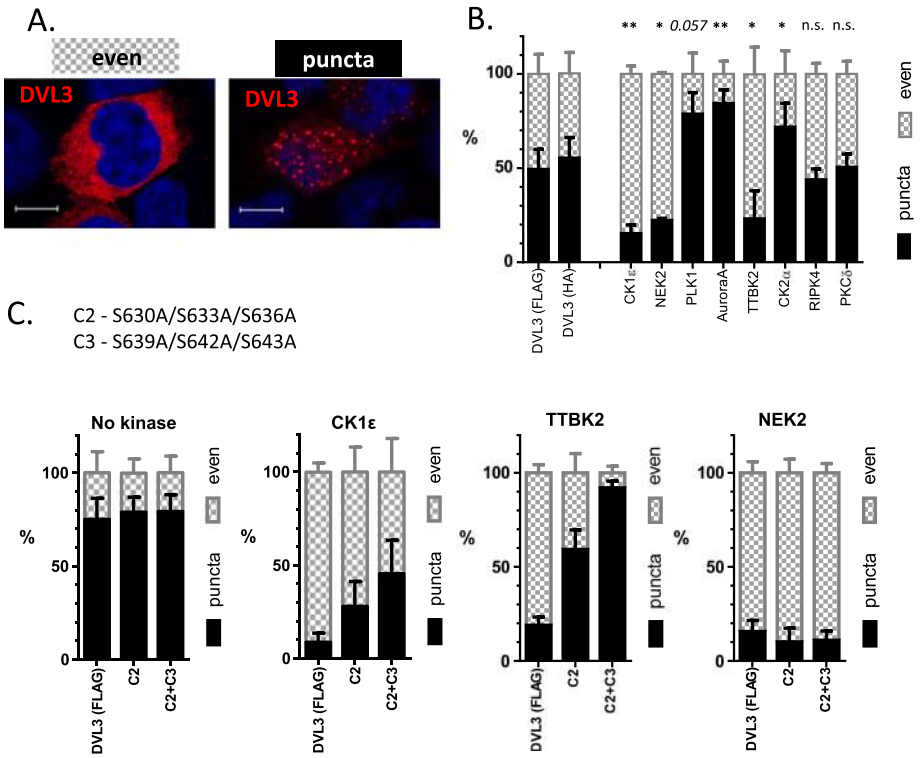


Fig. 10 Phosphorylation of S630-S643 promotes even localization of DVL3. **a**: HEK293 cells were transfected in the indicated combinations and the subcellular localization of DVL3 was assessed by immunocytochemistry. DVL3 was localized in two typical patterns – either in cytoplasmic puncta or evenly dispersed in the cytoplasm (upper panel). Scale bar, 7.5 μm. **b**: The effects of individual kinases on DVL3 localization is shown in the bottom panel (HA-DVL3 was used for PLK1 and TTBK2, FLAG-DVL3 for the rest of kinases). **c**: Phosphorylation patterns associated with the even localization of DVL3 were analyzed by mutation of cluster of serine residues to alanine. Statistical data represent mean + SD from three independent experiments ($N = 3\times 200$ cells). Statistical significance was confirmed by the comparison of the corresponding control (DVL3-FLAG or DVL3-HA without kinase) and DVL3 with individual kinases by One-way ANOVA and Tukey's post test (* $p < 0.05$, ** $p < 0.01$, n.s. - not significant)

experience fast exchange perturbation (Fig. 11c). Several other peaks (e.g. F277 and K301) that are largely perturbed in the course of the titration in wildtype remain unaffected in mutant (Fig. 11c). The qualitative analysis of the titration experiments indicates that the phosphomimicking S281E/S298E PDZ is insufficient in peptide binding as compared to wt PDZ (Fig. 11c). This suggests that phosphorylation of S263 and S280 (corresponding to S281 and S298 in DVL2) attenuates the capacity of PDZ domain to bind the C-terminus of DVL3 and therefore leads to open conformation of DVL3.

To directly test whether NEK2 indeed induces open conformation of DVL3 we employed the DVL3 FRET sensors [24]. Specifically, the ECFP-DVL3 FlAsH III sensor (schematized in Fig. 11d) was transfected in HEK293 cells either alone or in the presence of wild type (WT) and kinase dead (KD) NEK2 and subsequently, the intramolecular FRET efficiency was quantified. As shown in Fig. 11e, only WT (but not KD) NEK2 was able to lower the intramolecular FRET efficiency that depends on the proximity of the ECFP and FlAsH (F) tag. This result suggests that the kinase activity of NEK2 (and not the presence of NEK2 itself as demonstrated by the KD variant) promotes the open conformation of DVL3. In summary with other data, we propose that NEK2-induced phosphorylation of S263 and S280 in human DVL3 prevents the interaction between DVL's C-terminus and PDZ domain, thus keeping DVL in the open conformation and more evenly distributed.

Discussion

Our study provides the first comprehensive description of DVL3 phosphorylation by most of the described DVL S/T kinases. Given the high sequence conservation of DVL3 and other DVL proteins it is an important reference point for the interpretation of published data as well as a benchmark for forthcoming studies focused on the regulation of DVL (and other proteins) function by kinases and phosphatases.

In this study we have identified TTBK2 and Aurora A as novel DVL kinases. For TTBK2 we propose a possible negative role in Wnt/β-catenin signaling. TTBK2 is found to basal bodies or distal appendages, respectively. Indeed, DVL has been associated with centrosome and

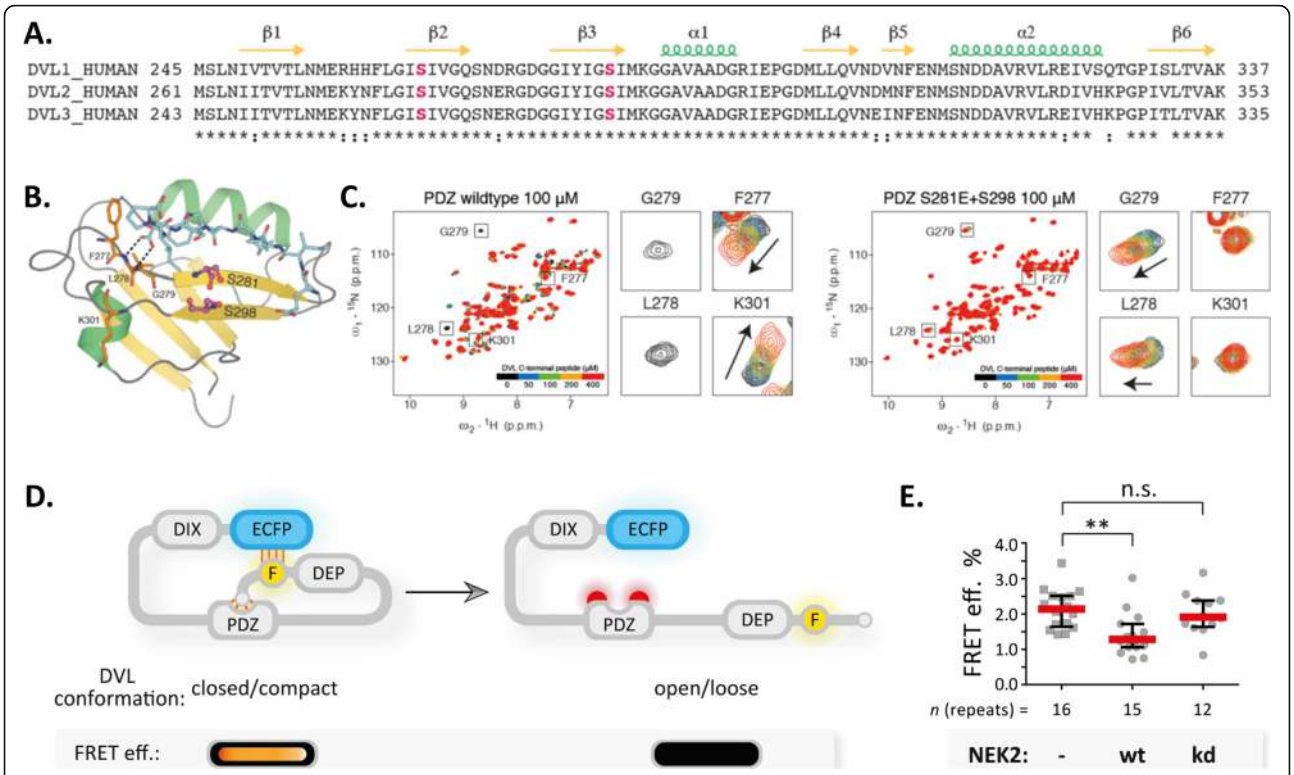


Fig. 11 NEK2 phosphorylation in PDZ domain promotes open conformation of DVL3. **a**: PDZ sequence alignment of the three human DVL isoforms. The conserved serine residues phosphorylated specifically by NEK2 are highlighted in magenta. **b**: Crystal structure of DVL2 PDZ domain bound to internal ligand (PDB: 3CC0). The two serines prone to NEK2 phosphorylation are shown in magenta, the bound peptide in cyan, and PDZ residues of the binding loop or close to the binding loop in orange. **c**: NMR titrations of DVL3 C-terminal peptide to DVL2 wildtype PDZ or S281E/S298E phosphomimicking mutant. Magnified insets show the qualitative differences in peptide binding for PDZ residues annotated in (**b**). **d**: Schematic depiction of the FRET efficiency experiment of the ECFP DVL3 FlAsH III construct. In the default conformation, the DVL molecule is less phosphorylated and closed/compact, which is reflected in the close proximity of the FRET pair (ECFP and FlAsH (F) tags) – leading to the high intramolecular FRET efficiency (depicted as orange dash lines). In the presence of the active kinase (e.g. NEK2), DVL is phosphorylated in the PDZ domain, which promotes the open/loose conformation by the disruption of the C-terminus/PDZ domain interaction. This is reflected in the high proximity of ECFP and F tags thus leading to the low intramolecular FRET efficiency (no orange dash lines). **e**: HEK293 cells were transfected in the indicated combinations and the intramolecular FRET efficiency of the ECFP-DVL3 FlAsH III construct (schematically depicted in D) was measured. The data represent median ± interquartile range from three independent experiments (numbers of analyzed cells are indicated below). Statistical significance was confirmed by One-way ANOVA and Tukey's post test (** $p < 0.01$, n.s. - not significant). FRET eff. Stands for Förster-Resonance-Energy-Transfer efficiency, ECFP for Enhanced Cyan Fluorescent Protein, F for FlAsH tag, and kd for kinase dead

basal bodies [6, 10, 53]. However, it does not seem to localize to distal appendages of mother centriole, where TTBK2 is recruited by CEP164 to trigger ciliogenesis [34, 40, 42–45]. It remains to be determined what is the possible interplay between TTBK2 activity and DVL3 phosphorylation by Aurora A, NEK2 and PLK1 and how these kinases together regulate DVL localization and functions in the individual phases of the cell cycle.

The parallel application of several pipelines for sample preparation and data analysis, in particular the newly introduced site occupancy analysis, demonstrated the potential to detect phosphorylation events that are biologically relevant. This approach identified S268 as signature site for CK1ε, S280 as a signature site for NEK2 and S633–S643 cluster as a signature for CK1ε, NEK2 and TTBK2. All these phosphorylation changes have been validated by additional, more demanding proteomic pipelines and functional analysis.

The data collected by various approaches shown in Figs. 4, 5, 6, 7, provide a global view of DVL3 phosphorylation induced by kinases that were shown to control distinct functions of DVL. We can clearly observe codes of phosphorylated S/T in the structured domains – most typically in the PDZ domain at the fully conserved S263, S268, S280 and S311, and to a lesser extent also in the DIX and DEP domains. On the other hand, in the largely intrinsically disordered regions (IDRs) between the domains, we have detected conserved (see Fig. 6) multiphosphorylated sequences with specific phosphorylation patterns. Most intrinsically disordered proteins or regions are very S/T rich because these aa (together with A, R, G, Q, P, E and K) are

disorder-promoting. IDR-rich proteins (such as DVL) are essential components of multiple signal transduction pathways [54] and complex multiphosphorylation as shown for DVL can represent a shared and universal mechanism for the regulation of their function.

It is behind the scope of this study to test the functional consequences of all phosphorylation events. However, several interesting candidates for further analysis have been identified. For example, a striking phosphorylation pattern has been observed for two highly conserved regions between DIX and DEP domains – corresponding to S112–S140 and S188–S197. For the second cluster we were able to confirm the proteomic data by a phosphorylation-specific antibody targeted against pS192-DVL3. These sites showed high level of basal phosphorylation in the absence of any exogenously co-expressed kinase. Interestingly, expression of CK1ε (and to a lesser extent TTBK2 and PLK1) dramatically reduced phosphorylation of these motifs. This suggests that there is a so far unidentified endogenous kinase that very efficiently phosphorylates these regions, and (ii) that the binding and/or phosphorylation by CK1ε and TTBK2 interferes with this process or perhaps promotes removal of these phospho-moieties by activation of specific phosphatase(s). So far, only protein phosphatase 2A has been shown to have a positive function in the Wnt/β-catenin signaling upstream of DVL [55, 56], which makes it an ideal candidate for such function.

Our functional analysis of PTMs that control subcellular localization of DVL3 showed that phosphorylation in the domains (namely PDZ) and in the IDRs can lead to the same outcome. CK1ε, NEK2 and TTBK2, were the only kinases that could trigger the even distribution of DVL3 in the cytoplasm, in line with earlier reports [10, 12, 13] [20]. Interestingly, phosphorylation of the cluster S630–S643 was essential for the ability of CK1ε and TTBK2 (but not of NEK2) to trigger even localization of DVL3. It appears that NEK2 uses a different mechanism dependent on the phosphorylation of S263 and S280 in the PDZ domain. Therefore, DVL C-terminus cannot bind to PDZ, which in turn stabilizes DVL3 in the open conformation. Importantly, despite the fact that the open conformation of DVL3 correlates with the even subcellular localization and more efficient recruitment to FZD receptors [24], it does not necessarily translate into the activation of the Wnt/β-catenin signaling. Our data are rather compatible with the hypothesis that phosphorylation opens an auto-inhibited DVL conformation and then the kinase-specific barcode dictates the biological outcome. Phosphorylated S268 and S311 can be part of such barcode required for CK1ε-induced activation of the Wnt/β-catenin signaling [11, 13] whereas phosphorylated S263 and S280 are barcode components essential for the NEK2-controlled functions of DVL3 in the centrosome [10].

Conclusions

In summary, our study identified unique DVL3 phosphorylation barcodes associated with individual DVL kinases and DVL functions. Our data represent an important reference point and a toolbox for further analysis of DVL as exemplified by the functional analysis of the phosphorylation in the PDZ domain. From the more general point of view, our data pinpoint the importance of functional synergy between phosphorylation in the structured domains and in the unstructured IDRs that together dictate the biological outcome.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12964-019-0470-z.

Additional file 1: Figure S1. SDS-PAGE gels from three independent experiments. FLAG-DVL3 was overexpressed in HEK293, with or without the studied kinase, immunoprecipitated using the anti-FLAG antibody, separated on SDS-PAGE and stained with Coomassie Brilliant Blue. In red boxes are parts that were cut out.

Additional file 2: Figure S2. Phosphorylated clusters map of DVL3. All identified phosphorylated clusters obtained from the pipeline #3 are visualized as a heatmap. Mean absolute intensities of phosphorylation clusters in the control (DVL3 without exogenous kinase; six replicates) are expressed in the shades of blue. Numbers indicate decadic logarithm of the mean of 6 control samples. Nine columns represent heat map of relative change of phosphorylated peptide intensities (in log10 scale) obtained for individual kinases (relative to control). Numbers in the heat map fields (0, 1, 2, 3) indicate the number of experimental replicates with the positive identification of the given phosphorylated site.

Additional file 3: Figure S3. Cluster analysis of the individual phosphorylated sites/clusters and individual kinases. Mean phosphorylated peptide intensities (in log10 scale) standardized to control are used in the circular plot. Peptides expressed more intensively compared to control are in the shades of green, whereas less expressed peptides are in red. Dendrograms both for the kinases and the phosphosites are shown.

Additional file 4: Figure S4. NMR comparison of DVL2 PDZ wild-type and phosphorylation-mimicking mutant S281E + S298E. A. 1H, 15 N HSQC spectra of PDZ wild-type (blue) and phosphorylation-mimicking mutant S281E + S298E (red). All amide frequencies have been assigned unambiguously using 4D spectra. The e cross-peaks of wild-type and corresponding glutamic acids cross-peaks of the mutant are marked. B. Weighted chemical shift differences between PDZ wild-type and S281E + S298E mutant versus the protein sequence.

Additional file 5: Table S1. Intensities of the individual phosphorylated sites (raw data). Mean phosphorylated peptide intensities (in log10 scale) obtained for individual sample from direct (pipeline #1) and enriched analysis processed (pipeline #2) are expressed in the shades of blue. ND corresponds to not detected signals or signals with intensity below 1×10^6.

Additional file 6: Table S2. Sequence coverage of DVL3 in the individual experiments. Sequence coverage of DVL3 obtained for individual kinases in each replicate, in individual sample (sum of all three replicates) and throughout the experiment.

Abbreviations

AB: Ammonium bicarbonate; CK: Casein kinase; DAPI: 4′,6-diamidino-2-phenylindole; DMEM: Dulbecco's modified eagle's medium; DTT: Dithiotreitol; DVL: Dishevelled protein in human; ECFP: Enhanced cyan fluorescent protein; EDT: 1,2-ethanedithiol; FCS: Fetal calf serum; FlAsH: Fluorescein arsenical hairpin binder; GSK: Glycogen synthase kinase; HBSS: Hank's balanced salt

solution; HCD: Higher-energy collisional dissociation; IAA: Iodoacetamide; IDR: Intrinsically disordered region; NEK2: Nima-related kinase 2; PEG: Polyethyleneglycol; PEI: Polyethylenimine; PFA: Paraformaldehyde; PKCδ: Protein kinase C δ; PLK1: Polo-like kinase 1; RIPK4: Receptor-interacting protein kinase 4; TCF/LEF: T-cell factor/lymphoid enhancer-binding factor; TTBK2: Tau tubulin kinase 2

Acknowledgements
We would like to thank Erich Nigg (Biozentrum, Basel) for sharing reagents, Lumir Krejci for access to DV Elite and Lucie Smyčková, Lenka Bryjová and Naďa Bílá for excellent assistance.

Authors' contributions
OB, MK, MM, PP,LC and MR performed all cell-based assays, JH performed FRET assays, KH prepared samples for MS analysis and processed the data, DP performed LC-MS/MS analyses, PO performed the statistical analyses, JK performed NMR experiments and analysis, VB, KH, DP, LC, KT and ZZ planned the experiments, interpreted the data and wrote the manuscript. All authors have approved the final version.

Funding
The work was supported by grant from the Czech Science Foundation (19-28347X) to V.B., K.T. and Z.Z, and grants from Czech Science Foundation (16-03269Y and 19-05244S to L. C.); Swiss National Science Foundation (IZ11Z0_166533 to L.C.); Follow up research fund from Federation of Biochemical and Biophysical Societies (FEBS) to L.C.; M.K. was supported by Czech Science Foundation grant GA18-17658S. Furthermore, this research was supported by project CEITEC 2020 (no. LQ1601) with financial contribution from the MEYS CR and National Programme for Sustainability II to K.H. O.B. was supported by funds from the Faculty of Medicine MU to junior researcher. M.K., P.P. and J.K. were supported by Brno Ph.D. Talent Scholarship – Funded by the Brno City Municipality. CIISB research infrastructure project LM2015043 funded by MEYS CR is gratefully acknowledged for the financial support of the LC-MS/MS measurements at the Proteomics Core Facility and NMR measurements at CEITEC Josef Dadok National NMR Centre.

Availability of data and materials
Proteomic data are available via PRIDE.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interest.

Author details
[1]CEITEC—Central European Institute of Technology, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic. [2]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Brno, Czech Republic. [3]Department of Experimental Biology, Faculty of Science, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic. [4]Department of Histology and Embryology, Faculty of Medicine, Masaryk University, Brno, Czech Republic. [5]Institute of Biostatistics and Analyses, Faculty of Medicine, Masaryk University, Brno, Czech Republic. [6]Department of Cytokinetics, Institute of Biophysics, Academy of Sciences of the Czech Republic, Brno, Czech Republic.

Received: 2 September 2019 Accepted: 22 October 2019
Published online: 23 December 2019

References

1. Nusse R, Clevers H. Wnt/β-catenin signaling, disease, and emerging therapeutic modalities. Cell. 2017;169(6):985–99.
2. Semenov MV, Habas R, Macdonald BT, He X. SnapShot: noncanonical Wnt signaling pathways. Cell. 2007;131(7):1378.
3. Schulte G, Bryja V. The frizzled family of unconventional G-protein-coupled receptors. Trends Pharmacol Sci. 2007;28(10):518–25.
4. Sharma M, Castro-Piedras I, Simmons GE, Pruitt K. Dishevelled: a masterful conductor of complex Wnt signals. Cell Signal. 2018;47(July):52–64.
5. Almuedo-Castillo M, Salo E, Adell T. Dishevelled is essential for neural connectivity and planar cell polarity in planarians. Proc Natl Acad Sci U S A. 2011;108(7):2813–8.
6. Park TJ, Mitchell BJ, Abitua PB, Kintner C, Wallingford JB. Dishevelled controls apical docking and planar polarization of basal bodies in ciliated epithelial cells. Nat Genet. 2008;40(7):871–9.
7. Lee KSH, Johmura Y, Yu L-R, Park J-E, Gao Y, Bang JK, et al. Identification of a novel Wnt5a-CK1ε-Dvl2-Plk1-mediated primary cilia disassembly pathway. EMBO J. 2012;31(14):3104–17.
8. Fumoto K, Kikuchi K, Gon H, Kikuchi A. Wnt5a signaling controls cytokinesis by correctly positioning ESCRT-III at the midbody. J Cell Sci. 2012;125(Pt 20):4822–32.
9. Kikuchi K, Niikura Y, Kitagawa K, Kikuchi A. Dishevelled, a Wnt signalling component, is involved in mitotic progression in cooperation with Plk1. EMBO J. 2010;29(20):3470–83.
10. Cervenka I, Valnohova J, Bernatik O, Harnos J, Radsetoulal M, Sedova K, et al. Dishevelled is a NEK2 kinase substrate controlling dynamics of centrosomal linker proteins. Proc Natl Acad Sci. 2016;113(33):9304–9.
11. Peters JM, McKay RM, McKay JP, Graff JM. Casein kinase I transduces Wnt signals. Nature. 1999;401(6751):345–50.
12. Bryja V, Schulte G, Rawal N, Grahn A, Arenas E. Wnt-5a induces Dishevelled phosphorylation and dopaminergic differentiation via a CK1-dependent mechanism. J Cell Sci. 2007;120(4):586–95.
13. Cong F, Schweizer L, Varmus H. Casein kinase Iepsilon modulates the signaling specificities of dishevelled. Mol Cell Biol. 2004;24(5):2000–11.
14. Bryja V, Schambony A, Čajánek L, Dominguez I, Arenas E, Schulte G. β-Arrestin and casein kinase 1/2 define distinct branches of non-canonical WNT signalling pathways. EMBO Rep. 2008;9(12):1244–50.
15. Willert K, Brink M, Wodarz A, Varmus H, Nusse R. Casein kinase 2 associates with and phosphorylates dishevelled. EMBO J. 1997;16(11):3089–96.
16. Bernatik O, Ganji RS, Dijksterhuis JP, Konik P, Cervenka I, Polonio T, et al. Sequential activation and inactivation of Dishevelled in the Wnt/beta-catenin pathway by casein kinases. J Biol Chem. 2011;286(12):10396–410.
17. Schertel C, Huang D, Bjorklund M, Bischof J, Yin D, Li R, et al. Systematic screening of a Drosophila ORF library in vivo uncovers Wnt/Wg pathway components. Dev Cell. 2013;25(2):207–19.
18. Kinoshita N, Iioka H, Miyakoshi A, Ueno N. PKC delta is essential for Dishevelled function in a noncanonical Wnt pathway that regulates Xenopus convergent extension movements. Genes Dev. 2003;17(13):1663–76.
19. Huang XD, McGann JC, Liu BY, Hannoush RN, Lill JR, Pham V, et al. Phosphorylation of dishevelled by protein kinase RIPK4 regulates Wnt signaling. Science. 2013;339(6126):1441–5.
20. Bernatik O, Sedova K, Schille C, Ganji RS, Cervenka I, Trantirek L, et al. Functional analysis of dishevelled-3 phosphorylation identifies distinct mechanisms driven by casein kinase 1 and frizzled5. J Biol Chem. 2014;289(34):23520–33.
21. Yanfeng WA, Berhane H, Mola M, Singh J, Jenny A, Mlodzik M. Functional dissection of phosphorylation of disheveled in Drosophila. Dev Biol. 2011;360(1):132–42.
22. Etheridge SL, Ray S, Li S, Hamblet NS, Lijam N, Tsang M, et al. Murine dishevelled 3 functions in redundant pathways with dishevelled 1 and 2 in normal cardiac outflow tract, cochlea, and neural tube development. PLoS Genet. 2008;4(11):e1000259.
23. Bernatík O, Šedová K, Schille C, Ganji RS, Červenka I, Trantírek L, et al. Functional analysis of dishevelled-3 phosphorylation identifies distinct mechanisms driven by casein kinase 1ε and Frizzled5. J Biol Chem. 2014;289(34):23520–33.
24. Harnoš J, Cañizal MCA, Jurásek M, Kumar J, Holler C, Schambony A, et al. Dishevelled-3 conformation dynamics analyzed by FRET-based biosensors reveals a key role of casein kinase 1. Nat Commun. 2019;10(1):1–18.
25. Lee KH, Johmura Y, Yu LR, Park JE, Gao Y, Bang JK, et al. Identification of a novel Wnt5a-CK1varepsilon-Dvl2-Plk1-mediated primary cilia disassembly pathway. EMBO J. 2012;31(14):3104–17.
26. Angers S, Thorpe CJ, Biechele TL, Goldenberg SJ, Zheng N, MacCoss MJ, et al. The KLHL12-Cullin-3 ubiquitin ligase negatively regulates the Wnt-beta-catenin pathway by targeting Dishevelled for degradation. Nat Cell Biol. 2006;8(4):348–57.
27. Foldynová-Trantírková S, Sekyrová P, Tmejová KK, Brumovská E, Bernatík OO, Blankenfeldt W, et al. Breast cancer-specific mutations in CK1epsilon inhibit

Wnt/beta-catenin and activate the Wnt/Rac1/JNK and NFAT pathways to decrease cell adhesion and promote cell migration. Breast Cancer Res. 2010; 12(3):1–14.
28. Fry AM, Meraldi P, Nigg EA. A centrosomal function for the human Nek2 protein kinase, a member of the NIMA family of cell cycle regulators. EMBO J. 1998;17(2):470–81.
29. Golsteyn RM, Schultz SJ, Bartek J, Ziemiecki A, Ried T, Nigg EA. Cell cycle analysis and chromosomal localization of human Plk1, a putative homologue of the mitotic kinases Drosophila polo and Saccharomyces cerevisiae Cdc5. J Cell Sci. 1994;107(Pt 6):1509–17.
30. Meraldi P, Honda R, Nigg EA. Aurora-a overexpression reveals tetraploidization as a major route to centrosome amplification in p53–/– cells. EMBO J. 2002;21(4):483–92.
31. Turowec JP, Duncan JS, French AC, Gyenis L, St. Denis NA, Vilk G, et al. Protein kinase CK2 is a constitutively active enzyme that promotes cell survival: strategies to identify CK2 substrates and manipulate its activity in mammalian cells. Methods Enzymol. 2010;484:471–93.
32. Kajimoto T, Sawamura S, Tohyama Y, Mori Y, Newton AC. Protein kinase C {delta}-specific activity reporter reveals agonist-evoked nuclear activity controlled by Src family of kinases. J Biol Chem. 2010;285(53):41896–910.
33. Bertrand MJM, Lippens S, Staes A, Gilbert B, Roelandt R, De Medts J, et al. cIAP1/2 Are Direct E3 Ligases Conjugating Diverse Types of Ubiquitin Chains to Receptor Interacting Proteins Kinases 1 to 4 (RIP1–4). Gartel AL, editor. PLoS One. 2011;6(9):e22356.
34. Cajánek L, Nigg E. a. Cep164 triggers ciliogenesis by recruiting tau tubulin kinase 2 to the mother centriole. Proc Natl Acad Sci U S A. 2014;111(28): E2841–50.
35. Yan X, Habedanck R, Nigg EA. A complex of two Centrosomal proteins, CAP350 and FOP, cooperates with EB1 in microtubule anchoring. Mol Biol Cell. 2006;17(2):634–44.
36. Hoffmann C, Gaietta G, Bünemann M, Adams SR, Oberdorff-Maass S, Behr B, et al. A FlAsH-based FRET approach to determine G protein-coupled receptor activation in living cells. Nat Methods. 2005;2(3):171–6.
37. Evangelidis T, Nerli S, Nováček J, Brereton AE, Karplus PA, Dotas RR, et al. Automated NMR resonance assignments and structure determination using a minimal set of 4D spectra. Nat Commun. 2018;9(1):384.
38. Stejskal K, Potěšil D, Zdráhal Z. Suppression of peptide sample losses in autosampler vials. J Proteome Res. 2013;12(6):3057–62.
39. Chaki M, Airik R, Ghosh AKK, Giles RHH, Chen R, Slaats GGG, et al. Exome capture reveals ZNF423 and CEP164 mutations, linking renal ciliopathies to DNA damage response signaling. Cell. 2012;150(3):533–48.
40. Goetz SC, Liem KF, Anderson KV. The spinocerebellar ataxia-associated gene tau tubulin kinase 2 controls the initiation of ciliogenesis. Cell. 2012;151(4): 847–58.
41. Tanos BE, Yang HJ, Soni R, Wang WJ, Macaluso FP, Asara JM, et al. Centriole distal appendages promote membrane docking, leading to cilia initiation. Genes Dev. 2013;27(2):163–8.
42. Oda T, Chiba S, Nagai T, Mizuno K. Binding to Cep164, but not EB1, is essential for centriolar localization of TTBK2 and its function in ciliogenesis. Genes Cells. 2014;19(12):927–40.
43. Xu Q, Zhang Y, Wei Q, Huang Y, Hu J, Ling K. Phosphatidylinositol phosphate kinase PIPKIγ and phosphatase INPP5E coordinate initiation of ciliogenesis. Nat Commun. 2016;7:10777.
44. Yang TT, Chong WM, Wang WJ, Mazo G, Tanos B, Chen Z, et al. Super-resolution architecture of mammalian centriole distal appendages reveals distinct blade and matrix functional components. Nat Commun. 2018;9(1):2023.
45. Huang N, Zhang D, Li F, Chai P, Wang S, Teng J, et al. M-phase Phosphoprotein 9 regulates ciliogenesis by modulating CP110-CEP97 complex localization at the mother centriole. Nat Commun. 2018;9(1):4511.
46. Bryja V, Červenka I, Čajánek L. The connections of Wnt pathway components with cell cycle and centrosome: side effects or a hidden logic? Crit Rev Biochem Mol Biol. 2017;52(6):614–37.
47. Paclíková P, Bernatík O, Radaszkiewicz TW, Bryja V. The N-Terminal Part of the Dishevelled DEP Domain Is Required for Wnt/β-Catenin Signaling in Mammalian Cells. Mol Cell Biol. 2017;37(18).
48. Kishida M, Hino SS, Michiue T, Yamamoto H, Kishida S, Fukui A, et al. Synergistic activation of the Wnt signaling pathway by Dvl and casein kinase Iepsilon. J Biol Chem. 2001;276(35):33147–55.
49. Schwarz-Romond T, Fiedler M, Shibata N, Butler PJ, Kikuchi A, Higuchi Y, et al. The DIX domain of Dishevelled confers Wnt signaling by dynamic polymerization. Nat Struct Mol Biol. 2007;14(6):484–92.
50. Lee HJ, Shi DL, Zheng JJ. Conformational change of Dishevelled plays a key regulatory role in the Wnt signaling pathways. Elife. 2015;4:e08142.
51. Gammons MV, Rutherford TJ, Steinhart Z, Angers S, Bienz M. Essential role of the Dishevelled DEP domain in a Wnt-dependent human-cell-based complementation assay. J Cell Sci. 2016;129(20):3892–902.
52. Zhang Y, Appleton BA, Wiesmann C, Lau T, Costa M, Hannoush RN, et al. Inhibition of Wnt signaling by Dishevelled PDZ peptides. Nat Chem Biol. 2009;5(4):217–9.
53. Gupta GD, Coyaud É, Gonçalves J, Mojarad BA, Liu Y, Wu Q, et al. A dynamic protein interaction landscape of the human centrosome-cilium Interface. Cell. 2015;163(6):1483–99.
54. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nat Rev Mol Cell Biol. 2015;16(1):18–29.
55. Yang J, Wu J, Tan C, Klein PS. PP2A:B56epsilon is required for Wnt/beta-catenin signaling during embryonic development. Development. 2003; 130(23):5569–78.
56. Ratcliffe MJ, Itoh K, Sokol SY. A positive role for the PP2A catalytic subunit in Wnt signal transduction. J Biol Chem. 2000;275(46):35680–3.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

Article

SOURCE DATA | TRANSPARENT PROCESS | OPEN ACCESS | CHECK FOR UPDATES

THE EMBO JOURNAL

Carboxy-terminal polyglutamylation regulates signaling and phase separation of the Dishevelled protein

Marek Kravec[1], Ondrej Šedo[2], Jana Nedvědová[3,4], Miroslav Micka[1,2], Marie Šulcová[1], Nikodém Zezula[1], Kristína Gömöryová[1], David Potěšil[2], Ranjani Sri Ganji[2], Sara Bologna[2], Igor Červenka[1], Zbyněk Zdráhal[2], Jakub Harnoš[1], Konstantinos Tripsianes[2], Carsten Janke[5,6], Cyril Bařinka[3] & Vítězslav Bryja[1]✉

Abstract

Polyglutamylation is a reversible posttranslational modification that is catalyzed by enzymes of the tubulin tyrosine ligase-like (TTLL) family. Here, we found that TTLL11 generates a previously unknown type of polyglutamylation that is initiated by the addition of a glutamate residue to the free C-terminal carboxyl group of a substrate protein. TTLL11 efficiently polyglutamylates the Wnt signaling protein Dishevelled 3 (DVL3), thereby changing the interactome of DVL3. Polyglutamylation increases the capacity of DVL3 to get phosphorylated, to undergo phase separation, and to act in the noncanonical Wnt pathway. Both carboxy-terminal polyglutamylation and the resulting reduction in phase separation capacity of DVL3 can be reverted by the deglutamylating enzyme CCP6, demonstrating a causal relationship between TTLL11-mediated polyglutamylation and phase separation. Thus, C-terminal polyglutamylation represents a new type of post-translational modification, broadening the range of proteins that can be modified by polyglutamylation and providing the first evidence that polyglutamylation can modulate protein phase separation.

Keywords Dishevelled 3; Polyglutamylation; TTLL11; Noncanonical Wnt Signaling; Protein Condensates
Subject Categories Development; Post-translational Modifications & Proteolysis; Signal Transduction
https://doi.org/10.1038/s44318-024-00254-7
Received 13 November 2023; Revised 15 August 2024; Accepted 16 September 2024
Published online: 30 September 2024

Introduction

Polyglutamylation is a reversible posttranslational modification (PTM) initially discovered on tubulin (Edde et al, 1990) that is catalyzed by several members of the tubulin tyrosine ligase- like (TTLL) protein family (Janke et al, 2005). Humans possess 13 TTLL homologs of which 9 are active glutamylases. According to the current understanding, tubulin polyglutamylation is initiated by the formation of the branching point by the addition of glutamate onto a γ-carboxyl group of a glutamate residue within the substrate, followed by the addition of multiple glutamates thereby generating secondary glutamate chains of variable lengths. Individual TTLL polyglutamylases show catalytic preferences either for the initial branching or the elongation of glutamate chains (Mahalingan et al, 2020; van Dijk et al, 2007). Moreover, individual enzymes also show preferences for different substrates such as α- or β-tubulin (van Dijk et al, 2007), or other, non-tubulin proteins out of which only a few have so far been identified (Regnard et al, 2000; van Dijk et al, 2008; Ye et al, 2018). Polyglutamylation is a reversible modification; deglutamylation is catalyzed by enzymes from the cytosolic carboxypeptidase (CCP) family (Kimura et al, 2010; Rogowski et al, 2010; Tort et al, 2014).

In our proteomic datasets, we discovered Dishevelled (DVL), a highly conserved regulatory protein in the Wnt signaling pathway, as a candidate binding partner of several TTLLs. Wnt pathways represent evolutionary conserved signaling modules that are required for multiple processes during the development and maintenance of homeostasis. The best-studied is the Wnt/β-catenin pathway (also referred to as "canonical") where DVL is necessary for membrane signalosome complex formation (Bilic et al, 2007; Kan et al, 2020) which triggers the disruption of cytoplasmic β-catenin destruction complex and subsequent stabilization of β-catenin. β-catenin is translocated into the nucleus where it activates T-cell factor/lymphoid enhancer factor (TCF/LEF) dependent transcription (Clevers and Nusse, 2012). Another important branch is the the Wnt/planar cell polarity (Wnt/PCP) pathway, which is responsible for polarized cellular orientation and cytoskeletal rearrangements during morphogenetic cell movements (Boutros et al, 1998; Butler and Wallingford, 2017; Devenport, 2014; Humphries and Mlodzik, 2018; Klingensmith et al, 1994; Koca et al, 2022; Wallingford et al, 2000). DVL acts as a central mediator in the

[1]Department of Experimental Biology, Faculty of Science, Masaryk University, Brno, Czech Republic. [2]Central European Institute of Technology (CEITEC), Brno, Czech Republic. [3]Institute of Biotechnology of the Czech Academy of Sciences, BIOCEV, Vestec, Czech Republic. [4]Department of Biochemistry, Faculty of Science, Charles University, Prague, Czech Republic. [5]Institut Curie, Université PSL, CNRS UMR3348, Orsay, France. [6]Université Paris-Saclay, CNRS UMR3348, Orsay, France. ✉E-mail: bryja@sci.muni.cz

© The Author(s)

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

transmission of Wnt signals, facilitating the assembly and localization of core PCP components at the cell membrane (Klingensmith et al, 1994; Wallingford et al, 2000). By interacting with various downstream effectors, such as RhoA, Rac1, and JNK (c-Jun N-terminal kinase) (Boutros et al, 1998; Harrison et al, 2020), DVL orchestrates the cytoskeletal rearrangements including cilia and centrosome positioning critical for the proper functioning of polarized cells (Hashimoto et al, 2010; Park et al, 2008). Three human DVL paralogs—DVL1, DVL2, and DVL3 act as common components for most, if not all, Wnt signaling branches (Gao and Chen, 2010; Wallingford and Habas, 2005; Wynshaw-Boris, 2012). DVL was proposed to act as a branching point between individual downstream signaling pathways. How this function is achieved is unclear, but it is expected that numerous DVL-interacting partners and modulation of DVL function by complex PTMs, primarily phosphorylation (Hanakova et al, 2019), have a key role. In addition to the role of DVL in Wnt pathways, it was shown to be localized in the basal bodies of cilia, affecting their positioning (Park et al, 2008), in the centrosome interacting with core centrosomal proteins (Cervenka et al, 2016; Park et al, 2008) or in the spindle poles and kinetochores during the mitosis (Kikuchi et al, 2010).

DVL and its interacting partners can form cytoplasmic puncta that have all the features typical for biomolecular condensates (Fiedler et al, 2011; Gammons et al, 2016; Schubert et al, 2022; Schwarz-Romond et al, 2007a; Schwarz-Romond et al, 2007b). In recent studies, endogenous DVL puncta were observed asymmetrically localized to the vegetal pole of sea urchin or sea star embryos as a part of an axis-defying event (Henson et al, 2021; Swartz et al, 2021b). In mammalian cells, endogenously tagged DVL2 formed condensates associated with centrosomal structures that were affected by cell cycle progression or Wnt signaling activity (Schubert et al, 2022). The puncta formation, mediated by the phase separation, has been shown to be regulated by phosphorylation and ubiquitination (Bernatik et al, 2014; Vamadevan et al, 2022). However, the dynamics of DVL phase transitions is not well-understood and was so-far studied mostly for phosphorylation mediated by casein kinase 1 (CK1) δ and ε (Harnos et al, 2019; Qi et al, 2017).

Following our observation of interactions between DVL and TTLL enzymes, we aimed at determining whether DVL is a substrate of polyglutamylation, and whether this is of biological importance. Testing all TTLL glutamylases, we found that TTLL11 can efficiently polyglutamylate DVL3. Surprisingly, the modification takes place via the addition of a polyglutamate chain at the α-carboxyl of C-terminal methionine of DVL3, and not, as so-far described as a branched (secondary) amino acid chain to the γ-carboxy group of the internal glutamate (Redeker et al, 1991). DVL3 polyglutamylation changed the interactome of DVL3, in cells lowered its propensity to undergo phase separation and increased the activity in the Wnt/PCP pathway in *Xenopus* embryos. Both C-terminal polyglutamylation and its impact on phase separation were reverted by the enzymatic activity of the CCP6 deglutamylase, suggesting that DVL3 polyglutamylation is a physiologically relevant regulatory mechanism.

Results

DVL proteins are polyglutamylated by TTLL11

We have recently identified several members of the TTLL protein family in the pull-downs of DVL3 (datasets partially published in (de Groot et al, 2014), see Dataset EV1), suggesting that TTLLs and DVL can interact. To assess whether any of the TTLLs can polyglutamylate DVL we overexpressed all 13 mouse TTLL paralogs tagged with EYFP together with Flag-tagged human DVL3 and performed western blotting (WB) analysis (Fig. 1A,B). We detected a change in the electrophoretic mobility of DVL3 when overexpressed with TTLL11, which indicated that TTLL11 might have modified DVL3. Moreover, analysis of polyglutamylation by an polyE antibody, which specifically recognizes polyglutamate peptides comprising at least 3 glutamates (Rogowski et al, 2010), revealed a prominent specific band in the presence of TTLL11 (in addition to signal corresponding to tubulin that served as a positive control) at the size of DVL3 (Fig. 1B). This signal was even more prominent when we used human TTLL11 (hTTLL11), indicating a higher efficiency of the human ortholog. We thus decided to use hTTLL11 in all subsequent experiments (unless specified otherwise). Importantly, co-expression of DVL3 with ligase-dead hTTLL11 (E531G) did not produce any polyglutamylation signal, revealing that the modification depends on the enzymatic activity of TTLL11 (Fig. 1C). The same experiments were performed with DVL2 and yielded identical results (Fig. EV1A,B). Next, we performed co-immunoprecipitation (co-IP) of DVL3 and TTLL11 (Fig. 1D) and could demonstrate that both proteins are not only present in one complex, but that DVL3 is also robustly polyglutamylated (Fig. 1D, WB: polyE). Similar results were obtained with other DVL paralogs DVL1 and DVL2 (Fig. EV1C). To show whether TTLL11 polyglutamylates DVL at endogenous levels, we performed IP of endogenous DVL3 and DVL2 in the presence and absence of exogenous TTLL11. As shown in Fig. 1E, both endogenous DVL3 and DVL2 were polyglutamylated (Fig. 1E). Together, this demonstrates that DVL proteins are substrates of TTLL11. In the further steps, we focused on DVL3 in order to describe the molecular details and significance of DVL3 polyglutamylation.

TTLL11 polyglutamylates DVL3 at the C-terminal methionine

To map the polyglutamylation site(s) of DVL3, we first performed domain mapping using a series of truncated versions of DVL3. All three human DVL paralogs have the same domain organization with an N-terminal DIX (Dishevelled, Axin), a central PDZ (Post-synaptic Density Protein-95, Disc Large Tumor Suppressor, and Zonula Occludens-1), and the most C-terminal DEP (Dishevelled, Egl-10, and Pleckstrin) domain. These conserved domains are connected and C-terminally extended by intrinsically disordered regions (Gao and Chen, 2010). As shown in Fig. 2A (for raw data see Fig. EV2A), the DVL3 C-terminus was essential for the polyglutamylation by TTLL11 as detected by the polyE antibody (IP Flag, WB PolyE). Posttranslational polyglutamylation is generated on glutamate residues of the primary peptide chain on tubulin (Edde et al, 1990) and also on the non-tubulin substrates like nucleosome assembly protein (NAP) (Regnard et al, 2000). Replacement of glutamate by the highly similar amino acid aspartate within the polyglutamylation site of tubulin completely abolished tubulin modification in *Tetrahymena*, an effect that can be explained by the structure of TTLL (Mahalingan et al, 2020; Thazhath et al, 2002). Since we have observed polyglutamylation of all DVL paralogs (Fig. 1), we mutated all glutamates (E) conserved in the C-terminus (corresponding to aa 496–716 of DVL3) of

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

 © The Author(s)

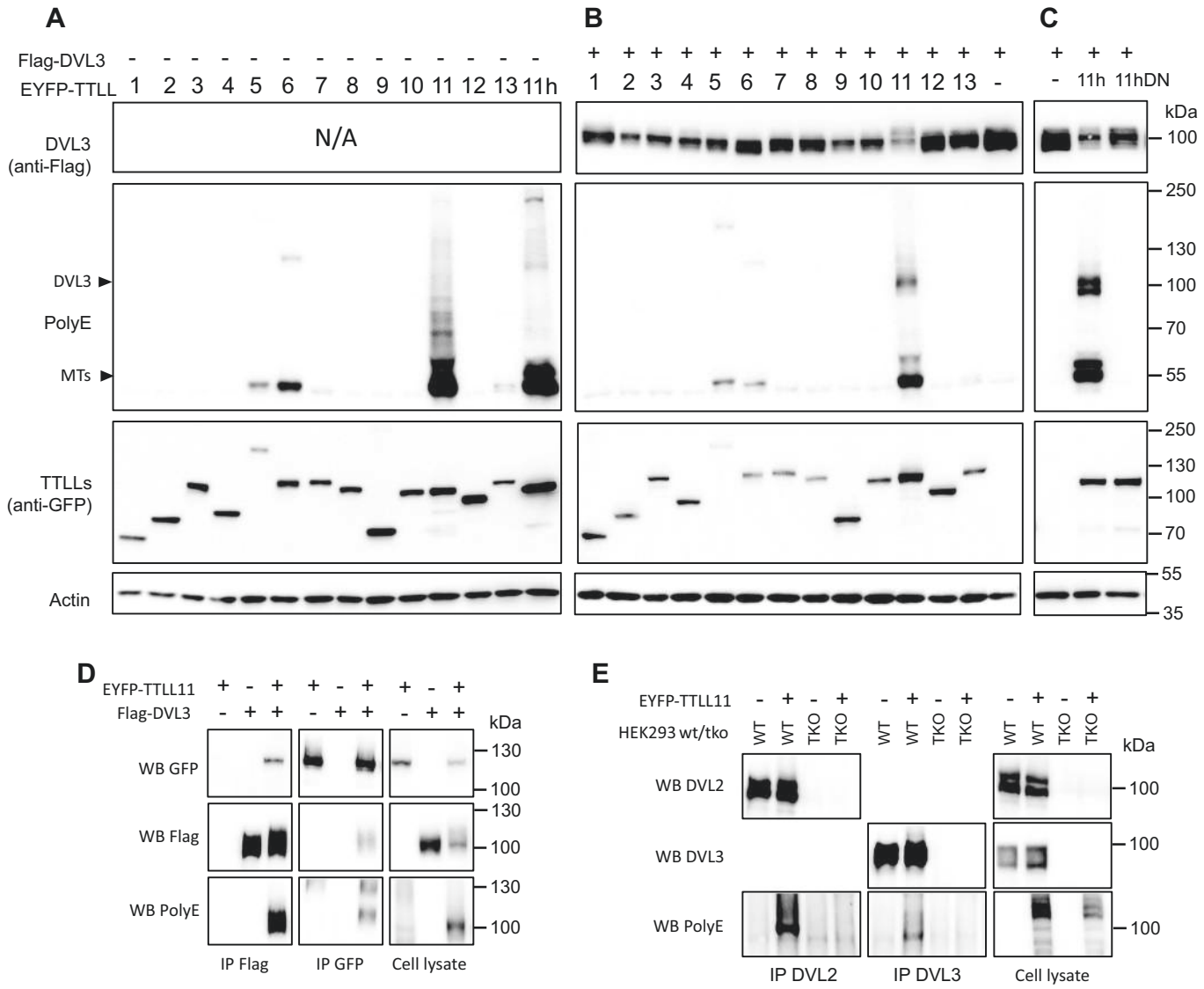


Figure 1. TTLL11 binds and polyglutamylates DVL3.

(**A–C**) HEK293T cells were transfected with constructs encoding murine EYFP-tagged TTLL1 – TTLL13 together with control plasmid (**A**) or DVL3 (**B**), or DVL3 with human TTLL11 and its inactive variant TTLL11 (E531G; DN) (**C**). The samples were subjected to WB analysis using polyglutamylation specific antibody—PolyE. Note the appearance of polyE-positive bands of DVL3 size in conditions where TTLL11 and DVL3 were co-expressed. The corresponding experiment with DVL2 is shown in Fig. EV1A,B. (**D**) Co-immunoprecipitation (co-IP) of Flag-DVL3 and TTLL11 overexpressed in HEK293T cells. TTLL11 is co-immunoprecipitated with DVL3; DVL3 in the pulldown was polyglutamylated when co-expressed with TTLL11 (IP Flag, WB PolyE). (**E**) IP of endogenous DVL3 and DVL2 from HEK293T cells in presence or absence of overexpressed TTLL11. Both endogenous DVL2 and DVL3 were polyglutamylated (WB PolyE). *DVL1/DVL2/DVL3* triple knockout (TKO) HEK293T cells served as a negative control. WB western blot, IP immunoprecipitation, MTs microtubules, WT wild type. Source data are available online for this figure.

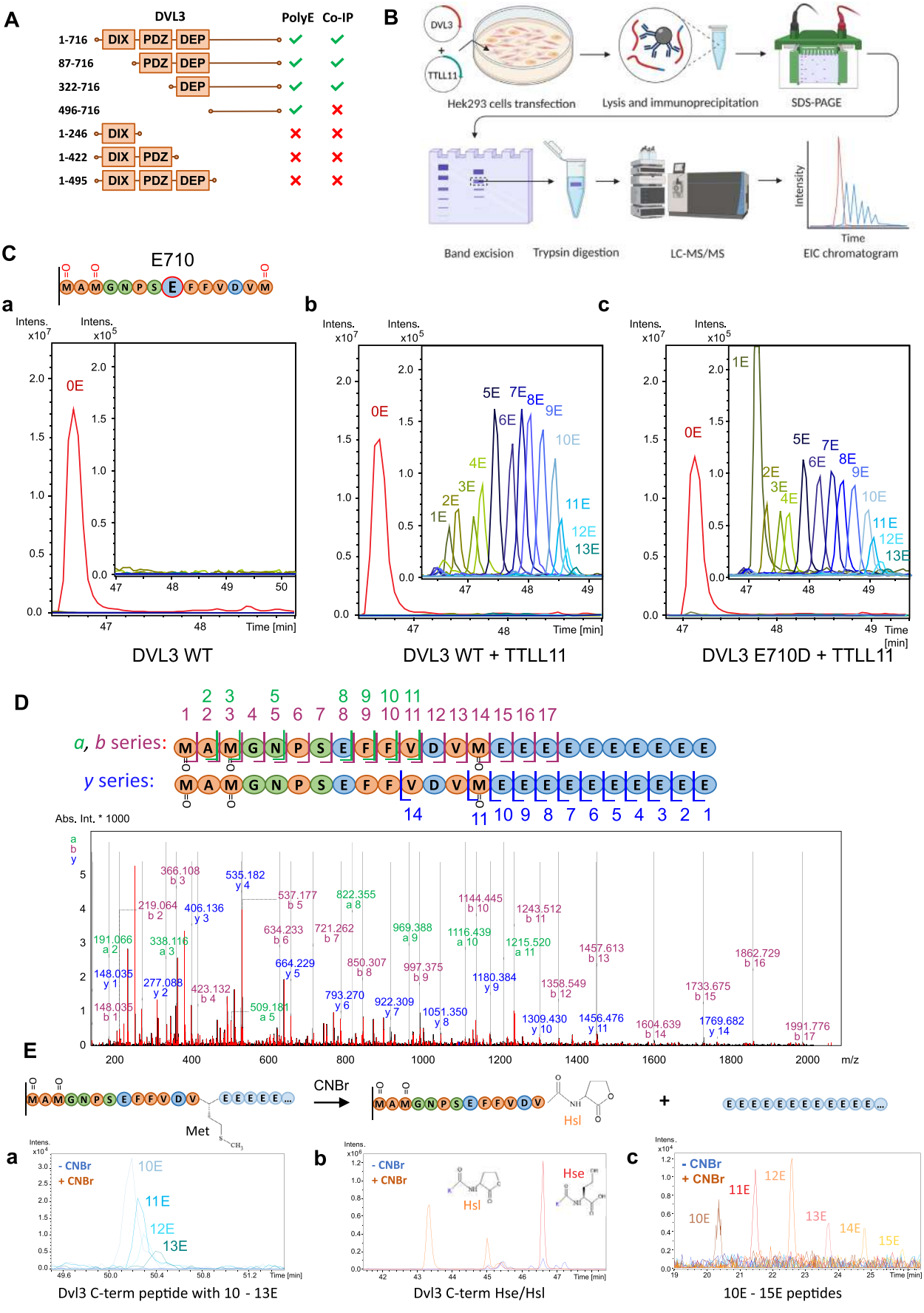
DVL1, DVL2 and DVL3 into aspartates (D). Subsequently, we compared polyglutamylation levels of wild-type (WT) and E571D/E604D/E693D/E710D DVL3 using the polyE antibody, and, strikingly, did not observe any differences (Fig. EV2B).

To identify the exact polyglutamylation site on DVL3, we used a mass spectrometry (MS) approach schematized in Fig. 2B. Analysis of tryptic digests of DVL3, which was co-expressed with TTLL11 in HEK293 and immunoprecipitated, by LC-MS/MS mapped polyglutamylation to the very last C-terminal DVL3 tryptic peptide—MAMGNPSEFFVDVM—with the most abundant peaks corresponding to peptides with an additional 7–9 glutamate residues (Fig. 2Ca,b). Of note, the peptide contains three methionine residues (M703, M705, and M716) that get partially oxidized during sample processing. In Fig. 2C, we present signals for the fully oxidized peptide; for the experiment where we performed the analysis of all oxidation variants, see Fig. EV2C,D.

We hypothesized that, similarly to tubulin and NAP, glutamylation of this peptide is realized by branching and further polyglutamate chain extension at the sole internal glutamate (E710). Interestingly, however, the E710D DVL3 mutant was

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

© The Author(s)



Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

 © The Author(s)

Figure 2. Modification site of DVL3 polyglutamylation and a novel character of the PTM.

(**A**) Domain mapping of DVL3 polyglutamylation. DVL3 truncation mutants were co-expressed with TTLL11 in HEK293T cells, DVL3 was immunoprecipitated, and interaction with TTLL11 and DVL3 polyglutamylation was detected by anti-GFP and PolyE antibody, respectively. Raw data are shown in Fig. EV2A. (**B**) Scheme of the sample preparation for mass spectrometry (MS)-based analysis of polyglutamylation. (**C**) MS analysis of DVL3 polyglutamylation. The C-terminal tryptic peptide of DVL3 (schematized) was found to be polyglutamylated by TTLL11. EIC chromatogram shows peaks of this peptide in DVL3 (**a**), DVL3 co-expressed with TTLL11 (**b**) and DVL3 E710D in the presence of TTLL11 (**c**). Polyglutamylated variants (blue lines) which are zoomed in a separate window. Only peaks corresponding to 3x M-ox peptide are shown, all peaks for 0–3x M-ox C-terminal peptides are shown in Fig. EV2C,D. (**D**) MS/MS fragmentation analysis of polyglutamylated DVL3 C-terminal peptide (3x M-ox variant). The peaks in a, b and y series observed are highlighted by green, purple, and blue color respectively. MS/MS spectra of control synthetic peptides with 10E modification at M716 or E710 in comparison to the polyglutamylated DVL3 immunoprecipitated from cells is shown in Appendix Fig. S1. The aa sequences for individual fragments are shown in the schematic representation above the chromatogram. (**Ea–c**) Confirmation of C-terminal modification by CNBr cleavage. EIC shows peaks corresponding to polyglutamylated DVL3 C-terminal peptide before (blue) or after CNBr treatment (orange). C-terminal homoserine (Hse), homoserine lactone (Hsl), and free glutamate chains (10E–15E) are detected only after CNBr treatment. Hse or Hsl with polyE chain were not detected. The data shown correspond to 2x M-ox peptide; other M-ox variants are shown in Appendix Fig. S2A–D.

polyglutamylated to the same extent as DVL3 WT (Fig. 2Cc). Detailed analyses of peptide fragmentation by MS/MS suggested the presence of a linear chain of glutamic acid residues after C-terminal M716 of the peptide, which was visible, especially in *b*- and *y*-ion series (Fig. 2D). Such a modification, i.e., the elongation of the primary amino acid chain, was unexpected because it has never been reported before. Comparison of the MS/MS spectra of the C-terminal tryptic peptide of in-cell polyglutamylated DVL3 and synthetic peptides mimicking polyglutamylation either via branching at the internal E710 or by C-terminal elongation (Appendix Fig. S1) further supported this assumption. Among the MS/MS fragments of in-cell polyglutamylated DVL3, we did not find any fragments characteristic of branched peptide but observed several fragments unique for the linear polyglutamylated peptide. Altogether, our data strongly support the notion that TTLL11 catalyzes a so-far unknown type of polyglutamylation that extends the primary peptide sequence of DVL3 by the addition of multiple glutamates beyond the C-terminal methionine residue.

In order to directly prove C-terminal polyglutamylation at M716, we immunoprecipitated DVL3 that was co-expressed with TTLL11 and performed cyanogen bromide (CNBr) cleavage of the mixture of DVL3 tryptic peptides. CNBr cleaves peptide bonds at the C-terminus of non-oxidized methionine residues, releasing peptides with C-terminal homoserine (Hse) or homoserine lactone (Hsl) (Fig. 2E). After CNBr cleavage, the peak series corresponding to polyglutamylated MAMGNPSEFFVDVM disappeared (Fig. 2Ea), and instead non-glutamylated peptides terminated either by Hse or Hsl appeared (Fig. 2Eb). In addition, after CNBr treatment, we could detect signals corresponding to free polyglutamate chains of 10–15 glutamate residues (Fig. 2Ec). This effect was the most obvious in the double oxidized peptide (Fig. 2E), but the other oxidized forms of the C-terminal peptide showed similar results with the notable exception of the fully oxidized peptide that could not be digested by CNBr and thus conveniently served as a negative control (Appendix Fig. S2A–D). These results independently confirmed that DVL3 is polyglutamylated by the extension of the C-terminal M716 and provided further evidence for the existence of this novel type of PTM catalyzed by TTLL11.

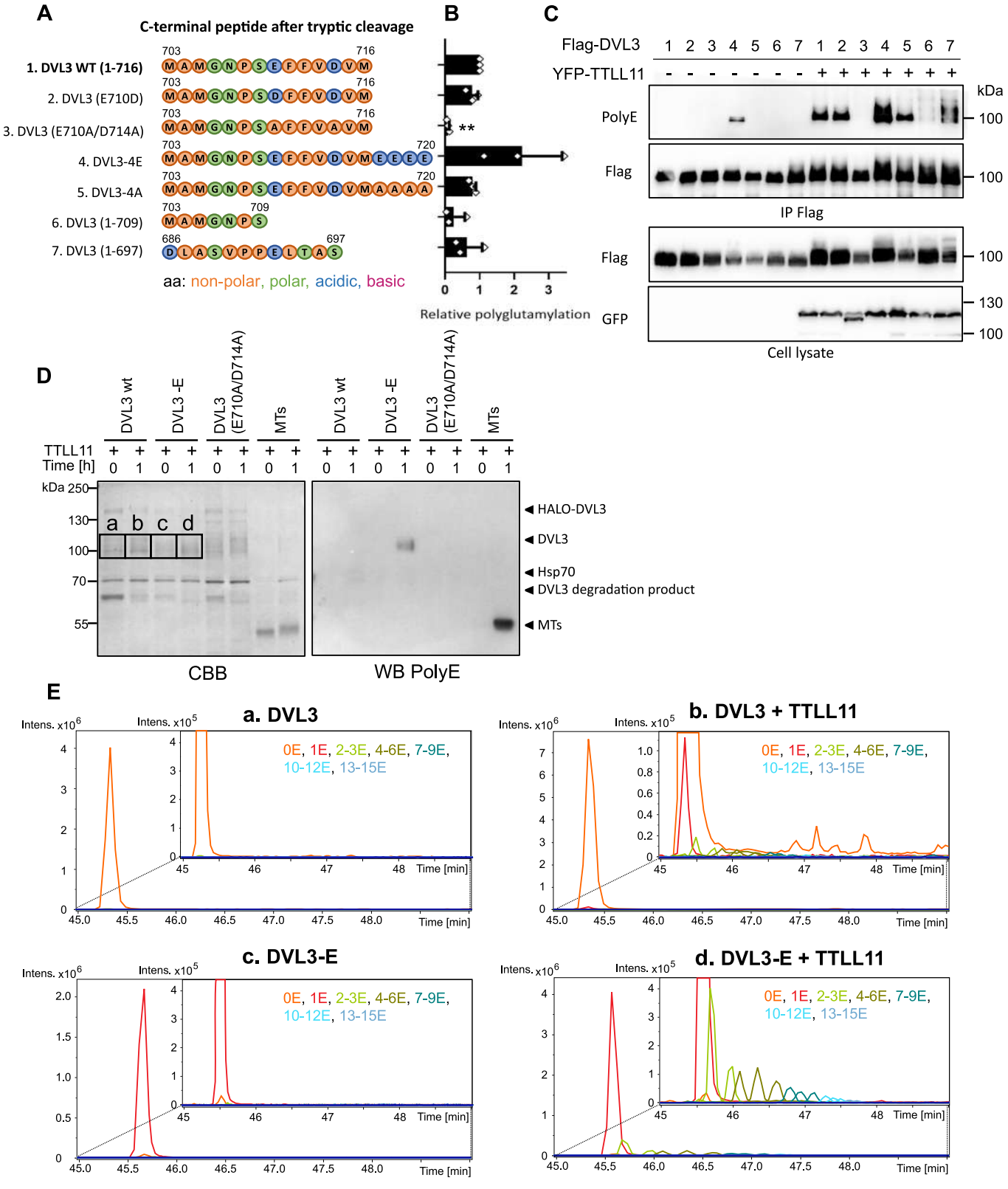
Sequence determinants of TTLL11-mediated DVL3 polyglutamylation

To determine molecular signatures of C-terminal polyglutamylation catalyzed by TTLL11, we generated a panel of DVL3 C-terminal mutants, where we (1) mutated the C-terminal glutamate residue 710 to aspartate (E710D); (2) mutated both acidic residues to alanines (E710A/D714A); (3) elongated the main polypeptide chain by the addition of four glutamate (4E) or four alanine (4A) residues; and (4) generated truncated variants lacking the C-terminal tail (DVL3 1–709; 1–697) (Fig. 3A). These mutants were then compared for their capacity to be modified by TTLL11. To this end, DVL3 was immunoprecipitated and polyglutamylation was analyzed by WB with the polyE antibody (Fig. 3B,C), as well as by MS-based detection (Fig. EV3A). Results from both experimental approaches were in good agreement (compare Figs. 3B and EV3A) and pinpointed several important facts. First, the modification is not strictly dependent on the presence of M716 or the exact C-terminal sequence as such because truncation mutants DVL3 1–697 and 1–709 as well as DVL3-4A were modified similarly to DVL3. Second, DVL3 E710A/D714A was not polyglutamylated at all and as such it could serve as a negative non-glutamylatable DVL3 control in subsequent functional experiments. Third, TTLL11 was clearly much more efficient in the polyglutamylation of DVL3-4E with the preexisting glutamate chain. To further study this effect, we generated DVL3 with only one extra glutamate (DVL3-E) at its C-terminus. As shown in Fig. EV3B already addition of one glutamate was sufficient to enhance polyglutamylation when compared to DVL3 WT.

In order to prove that TTLL11 directly polyglutamylates DVL3, we set up an in vitro polyglutamylation reaction using purified recombinant full-length DVL3 and TTLL11. We used DVL3 WT, DVL3-E, and the glutamylation-defective E710A/D714A mutant. Following in vitro incubation, the assay mixtures were analyzed by WB (Fig. 3D) and MS (Fig. 3Ea–d). Using WB, we could detect in vitro polyglutamylation by the polyE antibody only on DVL3-E. Interestingly, MS analysis not only confirmed polyglutamylation of DVL3-E but also detected monoglutamylated as well as polyglutamylated peptides in the DVL3-WT sample to a lower degree. This result suggests that TTLL11 itself is capable of DVL3 polyglutamylation but the ligation of the first glutamate can represent the rate-limiting step. It is possible that the in vitro assay with purified components is less efficient than polyglutamylation of DVL3 in cells, which could explain the difference in the number of added glutamates between these two assays. Notwithstanding this difference, our in vitro assay unambiguously demonstrates the direct and cofactor-independent modification of DVL3 by TTLL11.

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

© The Author(s)



Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

 © The Author(s)

Figure 3. Sequence determinants of TTLL11-mediated DVL3 polyglutamylation.

(A) Schematic representation of DVL3 mutants. **(B, C)** HEK293 cells were transfected with the indicated Flag-DVL3 mutants and EYFP-TTLL11. DVL3 variants were immunoprecipitated and the extent of polyglutamylation was analyzed by anti-PolyE antibody ($n = 3$ biological replicates). Relative polyglutamylation is represented as PolyE band intensity, normalized to total protein (Flag) signal, and shown as a fold change compared to Dvl3 WT polyglutamylation. Statistical significance was assessed using one-sample *t* test with theoretical median = 1; $^{**}P = 0.0012$. Graph represents mean + SD. **(D, E)** In vitro polyglutamylation of purified DVL3/DVL3-E by TTLL11. The level of polyglutamylation was analyzed by SDS-PAGE followed either by western blotting analysis using polyE antibody **(D)** or bands corresponding to DVL3 (indicated as **(a–d)** on Coomassie brilliant blue (CBB) stained gel) were excised and used for subsequent MS analysis. **(E)** EIC showing detection of in vitro polyglutamylated DVL3 C-terminal peptide in the individual conditions **(a–d)**. Source data are available online for this figure.

Polyglutamylation changes DVL3 interactome

Polyglutamylation adds a significant negative charge to target proteins and it might regulate protein-protein interactions of DVL3 as proposed for polyglutamylation of microtubules (MT) (Janke and Magiera, 2020). We thus decided to analyze how the interactome of DVL3 changes upon polyglutamylation. We over-expressed WT or glutamylation-defective (E710A/D714A) DVL3 in the presence and absence of TTLL11, performed a pulldown of DVL3, and identified interacting partners by tandem MS (MS/MS) (for schematics, see Fig. 4A). The differences in the interactome of DVL3 and DVL3 polyglutamylated by TTLL11 are shown in the Volcano plot in Fig. 4B. Numerous proteins showed a change in the binding to polyglutamylated DVL3 in the presence of TTLL11. To restrict these putative hits to those that are most likely caused by DVL3 polyglutamylation and not by interaction with TTLL11, we performed an additional comparison using REPRINT with integrated CRAPOME and SAINTexpress tools (Fig. 4C). This bioinformatic pipeline allows comparison of multiple samples and efficiently removes false positives (Choi et al, 2012; Mellacheruvu et al, 2013). Proteins significantly changed in REPRINT analysis are highlighted in black in Fig. 4B. We have focused on those REPRINT patterns (Fig. 4C) where TTLL11 decreased/increased abundance in the pulldown of wild type but not glutamylation-defective E710A/D714A DVL3. Such polyglutamylation-dependent pattern identified several candidates whose interaction with DVL3 can be reduced (SRSF9, DAP3, GRWD1) or increased (KATNAL2, RAB11FIP5) upon DVL3 polyglutamylation. These proteins were not reported as DVL-interacting partners before (Sharma et al, 2018), which was, however, not the case for other hits—casein kinase 1 (CK1) δ and ε (CSNK1D and CSNK1E). These enzymes are known as major kinases regulating the functions of DVL (Bernatik et al, 2011).

DVL3 polyglutamylation controls DVL3 activity in the Wnt/PCP but not Wnt/β-catenin signaling

We next aimed to determine the biological function of DVL polyglutamylation. DVL plays a key role in the activation of Wnt/β-catenin signaling, which induces T-cell factor/lymphoid enhancer factor (TCF/LEF)-dependent transcription. We thus first performed a series of experiments using TCF/LEF luciferase reporter (TopFlash assay) (schematized in Fig. EV4A). To test the impact of TTLL11 and DVL polyglutamylation on the Wnt/β-catenin signaling pathway. Interestingly, co-expression of TTLL11 with DVL1 (Fig. EV4Ba) or with DVL3 in combination with its activating kinase CK1ε (Kishida et al, 2001) (Fig. EV4Bb) efficiently reduced the TopFlash signal. Importantly, the luciferase signal decreased upon co-expression of both, WT as well as catalytically inactive TTLL11 (E531G); (Fig. EV4B). This suggests that the excess of TTLL11 inhibits Wnt/β-catenin signaling independently of its enzymatic activity and represents an artifact. Indeed, TTLL11 (E531G) was as potent in inhibiting the TopFlash reporter as TTLL11 WT, even under conditions that bypass the need for DVL by expressing the constitutively active mutant of the key Wnt receptor LRP6 (LRP6ΔN) (Tamai et al, 2004) (Fig. EV4Bc,d). The mechanism of how TTLL11 inhibits Wnt/β-catenin is unclear, but this fact thwarted our efforts to determine the molecular function of DVL3 polyglutamylation by simply overexpressing TTLL11.

As an alternative approach, we studied the properties of DVL3 WT, polyglutamylation-mimicking DVL3 with 4 and 12 C-terminally added glutamates (DVL3-4E and DVL3-12E), and glutamylation-defective DVL3 (E710A/D714A). First, we have tested to what extent DVL3-4E/12E and DVL3 E710A/D714A can transduce signals induced by Wnt3a. Rescue assays in *RNF43/ZNRF3/DVL1/DVL2/DVL3 penta* knockout cells (DVL PKO) (Paclíková et al, 2021) suggested that, albeit slightly less efficient than DVL3 WT, both DVL3-4E and -12E can well mediate Wnt3a-induced signaling (Fig. EV4Ca). Non-modifiable DVL3 E710A/D714A mutant showed reduced, but still potent ability to transduce Wnt3a-induced signaling. However, this DVL3 variant was repeatedly observed at lower levels in comparison to DVL3 WT or DVL3-12E (see Fig. EV4Cb), and as such, we do not think that the result reflects the lower capacity of DVL3 E710A/D714A to transduce Wnt3a signal. In line with our observation that polyglutamylation does not affect the function of DVL3 in Wnt/β-catenin signaling, co-expression of DVL3-activating kinase CK1ε with both DVL3 WT and DVL3-12E potentiated TCF/LEF-dependent transcription to similar levels (Fig. EV4D). In addition, Wnt ligands were unable to modify DVL3 interaction with TTLL11 or TTLL11-induced polyglutamylation of DVL3 (Fig. EV4E). We thus conclude that polyglutamylated DVL3 is proficient in the Wnt/β-catenin signaling and DVL3 polyglutamylation does not directly affect DVL role in the canonical Wnt pathway.

To address the possible role of DVL3 polyglutamylation in the noncanonical Wnt/PCP signaling, we have used the well-established model of the *Xenopus laevis* embryos (Hikasa and Sokol, 2013). Wnt/PCP in the *Xenopus* model controls the process of convergent extension (CE) during gastrulation and neurulation. The role of candidate proteins in CE can be assessed by the injection of mRNA or inhibitory morpholinos (MO) (for experimental schematics and scoring system, see Fig. 5A,B. Injection of the *Xenopus* (x) Dvl3 mRNA affected CE but, strikingly, the polyglutamylation-mimicking variant xDvl3-12E triggered CE defects much more efficiently (Fig. 5C). To strengthen this observation, we have analyzed also an earlier developmental

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

© The Author(s)

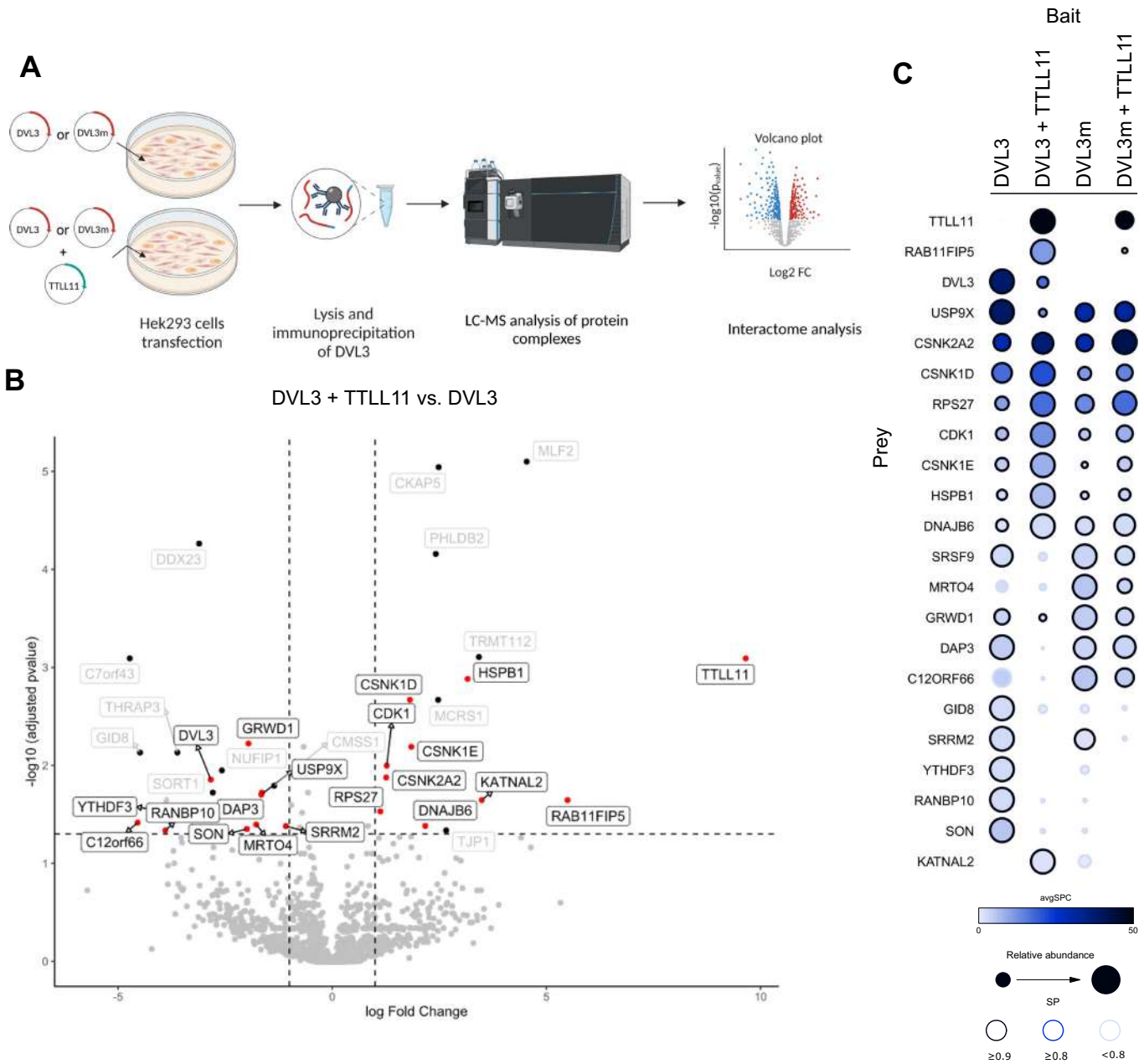


Figure 4. Change of DVL3 interacting partners upon its polyglutamylation.

(**A**) Scheme of sample preparation for MS analysis of DVL3 interactome upon co-expression with TTLL11. DVL3m = glytamylation defective DVL3 E710A/D714A. (**B**) Change of DVL3 interacting partners when overexpressed with TTLL11 was compared to overexpression of sole DVL3. DVL3 was immunopurified and interacting partners were analyzed by LC-MS in three biological replicates. The volcano plot shows detection of significantly upregulated and downregulated interacting partners using LIMMA statistical test—for more details, see "Methods" (FC ≥ 1, adjusted *P* value < 0.05). (**C**) Hits from LIMMA were further analyzed by REPRINT with integrated CRAPOME and SAINTexpress tools; average spectra (AvgSpec); saint probability (SP). The significantly changed proteins are highlighted in black in the Volcano plot (**B**).

event—blastopore closure (Fig. EV5A). The blastopore closes at stage 11.5; but gets delayed when CE movements mediated by PCP signaling are disrupted (Fig. EV5B). In line with data in Fig. 5C, polyglutamylated xDvl3-12E was able to delay the blastopore closure to a larger extent than xDvl3 WT (Fig. EV5C). The WB analysis in the embryo lysates confirmed that the levels of individual xDvl3 variants were comparable (Fig. EV5D). Interestingly, *xTTLL11* expression peaks exactly when CE movement during gastrulation starts (stages 8–11) in *X. laevis* embryos (Fig. EV5E,F). We have thus downregulated xTTLL11 by MOs targeting either its start codon (Fig. 5D) or splicing site (Fig. EV5G), which in both cases resulted in the defective CE.

The results in Fig. 5A–D suggest that TTLL11-mediated polyglutamylation of DVL is required for CE in *Xenopus*.

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

 © The Author(s)

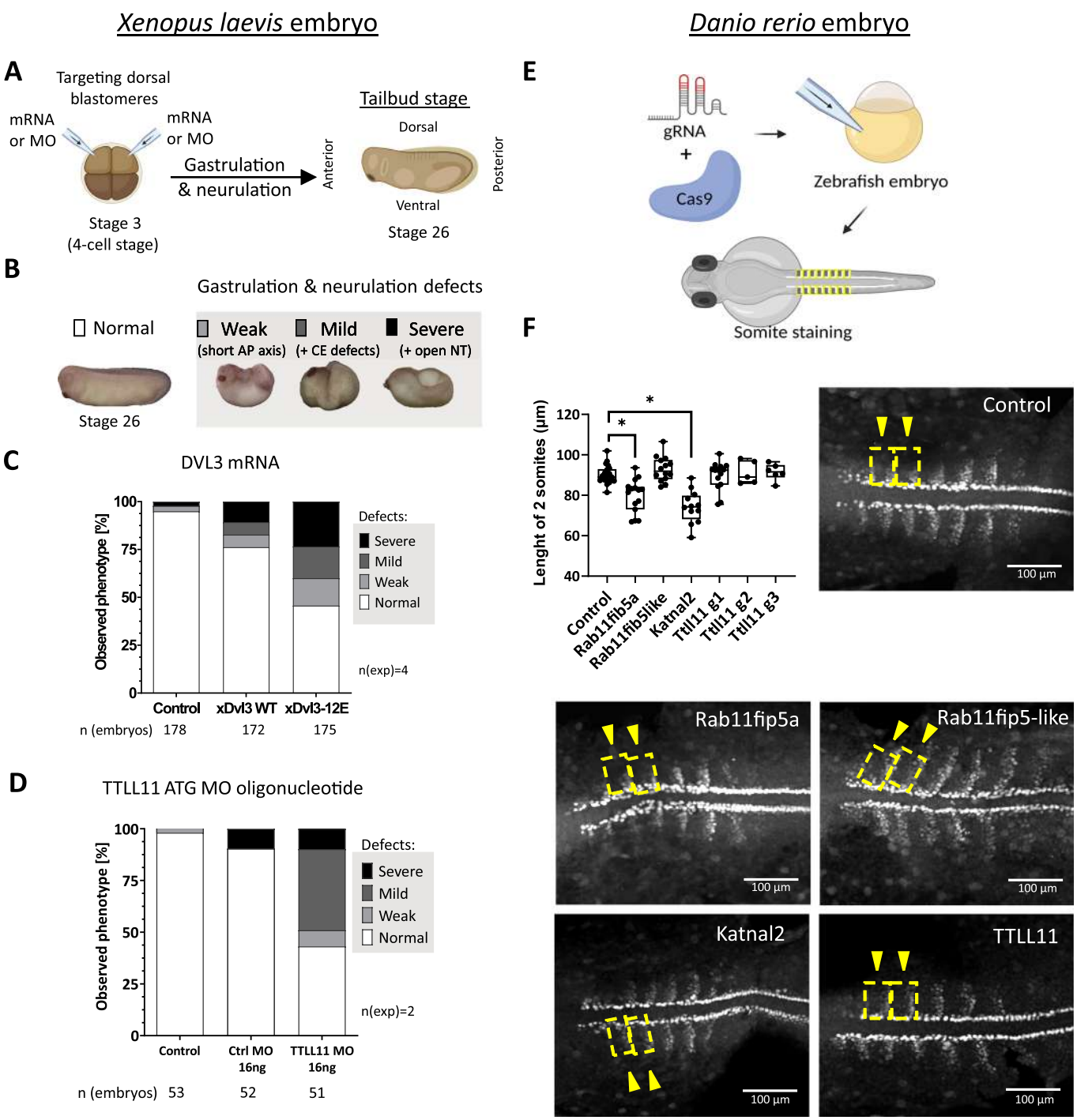
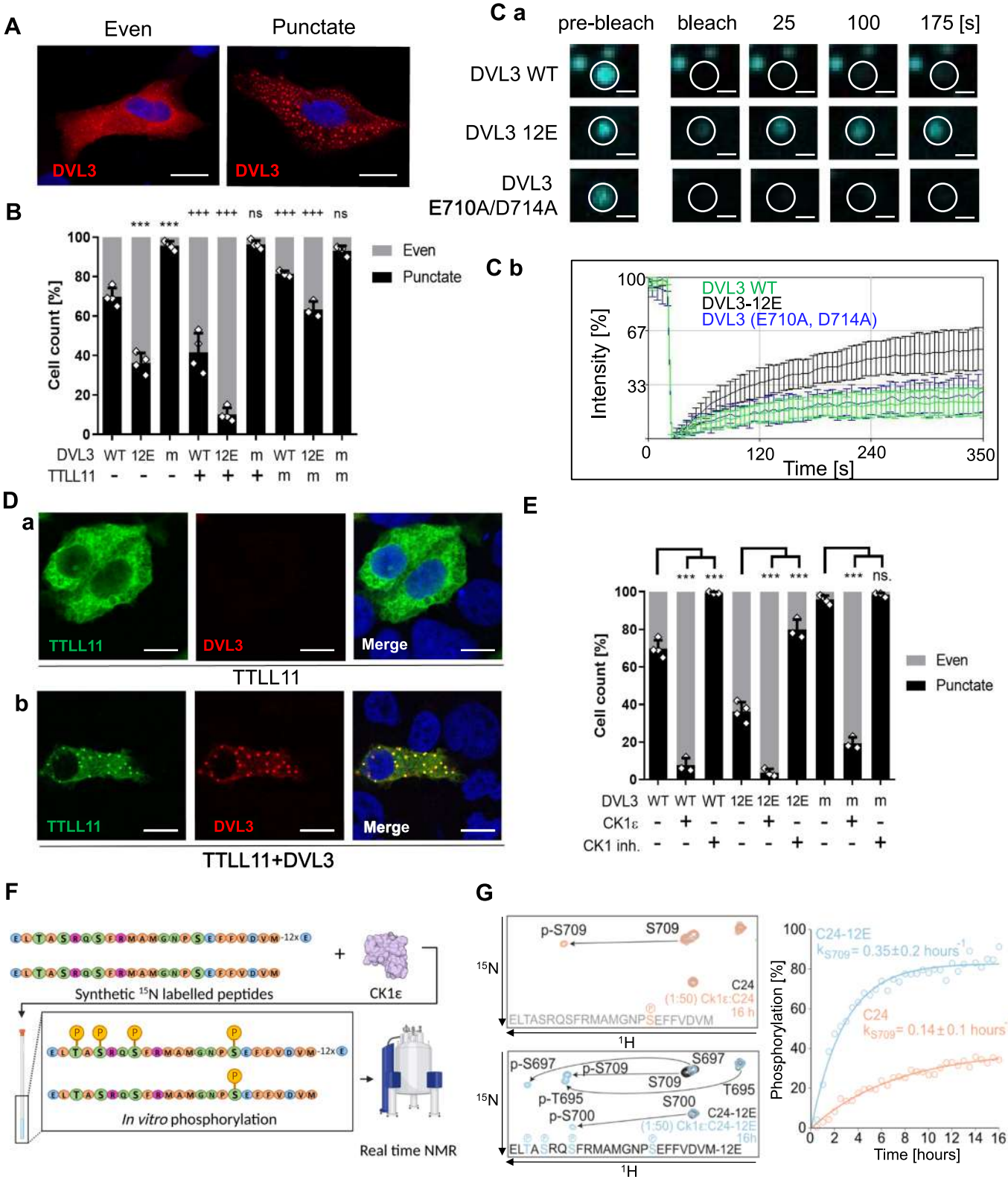


Figure 5. Role of DVL3 polyglutamylation in *Xenopus laevis* and *Danio rerio* embryonal development.

(**A**) Both dorsal blastomeres of 4-cell *Xenopus laevis* embryo were injected with xDvl3 mRNA or xTTLL11 morpholino (MO) and embryos were observed during gastrulation and neurulation at stage 26. (**B**) Phenotypes in *Xenopus* gastrulation and neurulation were assigned to four groups: no defects and weak (short AP axis), mild (+ convergent extension defects), and severe defects (+ open neural tube). (**C**) Effect of xDvl3 WT and xDvl3-12E modification mimicking mutant mRNA injection. (**D**) Effect of MO targeting xTTLL11 ATG start codon injection (for splicing MO see Fig. EV5G). (**E**) *Danio rerio* embryo was injected with Cas9 and gRNA targeting gene of interest. Embryos were fixed at 14.5 hpf and stained by α-MyoD1 antibody for somite visualization. (**F**) The length of the first 2 anterior somites was measured for control embryos and *Rab11fip5a*, *Rab11fip5-like*, *Katnal2*, and *TTLL11* KO embryos ($n > 12$). TTLL11 KO was additionally performed with two other gRNA TTLL11 g2 and g3 for verification of the result ($n > 5$). Box plots are shown as median (middle bar) with 25th and 75th percentiles and whiskers showing min to max values. Statistical analysis was performed by unpaired *t* test; $^*P < 0.05$. Source data are available online for this figure.

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

© The Author(s)



Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

© The Author(s)

Figure 6. Polyglutamylation of DVL3 alters its phase separation and promotes phosphorylation by Ck1ε.

(A) Overexpressed DVL3 in HEK293T cells. Scale bar = 10 µm. **(B)** Analysis of overexpressed DVL3 subcellular localization with its polyglutamylation variants in DVL1/DVL2/DVL3 triple KO HEK293 cells. The phenotype was assessed for >200 cells; $n \geq 3$. Mean cell counts with SD are indicated. Statistically significant difference from WT DVL3 (*), or from the same DVL3 variant in the absence of TTLL11 or inactive TTLL11m (E531G) (+) is indicated. ***/+++ represent $P < 0.001$; ns not significant. DVL3m = DVL3 E710A/D714A **(Ca)** FRAP analysis of ECFP-DVL3 WT, ECFP-DVL3-12E and ECFP-DVL3 E710A/D714A intracellular condensates. **(Cb)** Intensities of fluorescence summarized in the graph. The data represent average from 15 condensates in three biological replicates (5 droplets/replicate and error bars represent SD). Scale bar = 2 µm. Raw data are shown in Appendix Fig. S3C. **(D)** Overexpressed TTLL11 in the absence **(Da)** or in the presence **(Db)** of co-expressed DVL3 in HEK293T cells. Scale bar = 10 µm. **(E)** Analysis of subcellular localization of DVL3 variants upon co-expression or inhibition of CK1ε by 1 µM PF670462. Phenotype was assessed for >200 cells; $n \geq 3$ biological replicates. Mean values with SD are indicated. Statistically significant differences from the condition without CK1ε/CK1 inhibitor are indicated for each DVL3 variant (*). Experiments from 6B, 6E, and 7C were performed simultaneously, and control conditions are identical. For statistical analysis, see "Methods" and source data files. *** represent $P < 0.001$; ns not significant. **(F)** CK1ε in vitro phosphorylation of 24 DVL3-WT C-terminal aa (C24) or extended by 12E (C24-12E) peptides analyzed by NMR. **(G**, left) Overlay of HSQC spectra obtained before (gray and black for peptide C24 and C24-12E, respectively) and at the end of 16 h reaction (orange and blue for C24 and C24-12E, respectively). Arrows indicate chemical shift caused by phosphorylation. **(G**, right) Time course phosphorylation of C24 (orange) and C24-12E (blue) peptides at S709. The apparent rate constants were estimated from fitting to a single exponential function. Source data are available online for this figure.

Polyglutamylated DVL3 binds specifically KATNAL2 and RAB11-FIP5 (see Fig. 4). In order to test whether these downstream effectors also participate in CE, we have turned into zebrafish *Danio rerio* as a model. We have knocked out fish *Rab11fip5a*, *Rab11fip5-like*, and *Katnal2* genes as well as *Ttll11* by Crispr/Cas9 in zebrafish embryos (Fig. 5E) (Shah et al, 2015). Embryos upon deletion of *Rab11fip5a* and *Katnal2* showed body axis truncation phenotypes reminiscent of a defect in the PCP pathway (KO of *Wnt5b* served as the positive control, Fig. EV5H). This was further validated by quantitative analysis of the somite length at 14.5 hpf (Fig. 5F).

Altogether, these results suggested that polyglutamylated xDvl3 is more potent than xDvl3 WT in the activation of the Wnt/PCP pathway during CE movement. Consistently, downregulation of TTLL11 led to severe defects in gastrulation/neurulation on the *Xenopus* model, and KO of genes encoding interactors of the polyglutamylated DVL3 *Rab11fip5a* and *Katnal2* affected CE in *Danio rerio*. Collectively, this analysis supports the hypothesis that TTLL11-mediated polyglutamylation of DVL takes place during gastrulation to facilitate DVL-controlled morphogenetic processes such as CE movements.

Polyglutamylation controls the phase separation of DVL3

What can be the biochemical features of DVL3 affected by polyglutamylation? Since DVL3 can form biomolecular condensates (Sear, 2007), we have addressed next this possibility. Depending on the extent of phase separation, DVL3 is distributed either homogeneously in the cytoplasm, or forms phase-separated "puncta" (Fig. 6A).

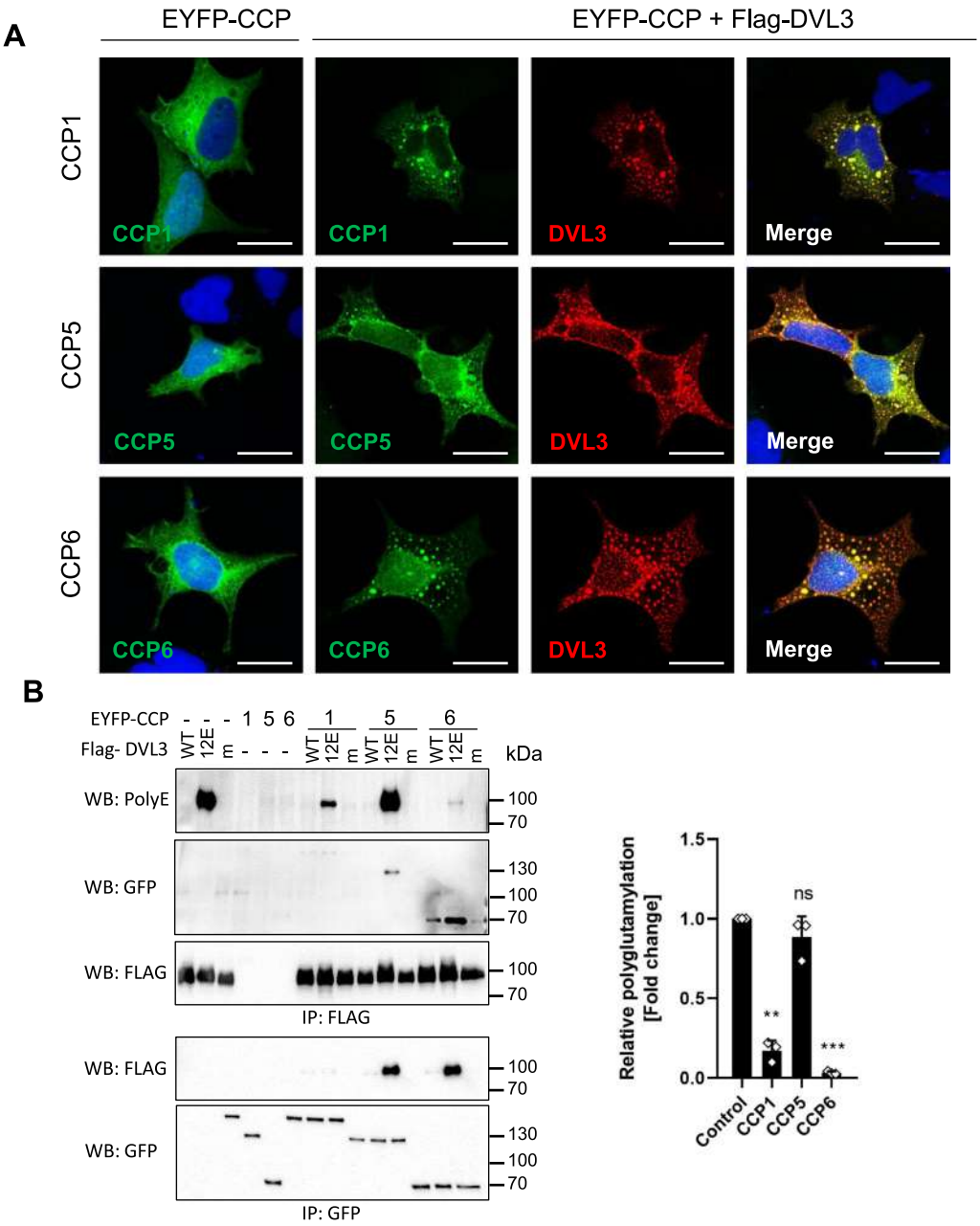
We thus overexpressed DVL3 WT, polyglutamylation-mimicking DVL3-12E, as well as the glutamylation-defective DVL3 E710A/D714A alone, or together with WT or ligase-dead TTLL11 (E531G) in HEK293 cells and analyzed the localization pattern of DVL3. To avoid interference with endogenous DVL proteins that were shown to affect this phenotype (Harnos et al, 2019) we have performed this experiment primarily in *DVL1/DVL2/DVL3* triple KO cells (Fig. 6B**;** the same experiment in WT HEK293 cells is provided for reference in Appendix Fig. S3A). Interestingly, polyglutamylation led to a decrease in DVL3 phase separation propensity, i.e., DVL3-12E was distributed more evenly in the cell cytoplasm compared to DVL3 WT (compare bars 1 and 2 in Fig. 6B). On the contrary, glutamylation-defective DVL3 E710A/D714A localized predominantly into puncta (bar 3, Fig. 6B). Co-expression of TTLL11 further promoted even localization of DVL3 WT and DVL3-12E but not of glutamylation-defective DVL3 E710A/D714A (Fig. 6B, compare bars 4–6 with bars 1–3), which suggested that this phenotypic difference is indeed because of polyglutamylation. In line with this assumption, ligase-dead TTLL11 (E531G) promoted puncta formation and phase separation of DVL3 (Fig. 6B, bars 7–9). However, the downregulation of *TTLL11* by siRNA did not show any effect (Appendix Fig. S3B, C), pointing towards functional redundancy of TTLL11 with other endogenously present TTLL proteins. To address the effect of polyglutamylation on DVL3 phase separation more directly, we analyzed the dynamics of ECFP-tagged wild-type DVL3, DVL3-12E, and DVL3 E710A/D714A using fluorescence recovery after photobleaching (FRAP) (Fig. 6C, raw data in Appendix Fig. S3E). We have selected cells where DVL3 forms puncta, bleached individual puncta of comparable size (Fig. 6Ca), and quantified the signal for 325 s. As shown in (Fig. 6Cb) polyglutamylated DVL3-12E showed much faster recovery of fluorescence than DVL3 WT or DVL3 E710A/D714A. This suggests that the turnover of polyglutamylated DVL3-12E between phase-separated puncta and other cytoplasmic pools of DVL3 is increased. This observation agrees with its reduced phase separation propensity. Of note, overexpressed TTLL11 localized in cells rather uniformly to the cytoplasm and filamentous structures (probably MT) (Fig. 6Da) when expressed alone. When we co-expressed DVL3 and focused on cells that still show punctate localization of DVL3 (~35% of cells, see Fig. 6B) TTLL11 co-localized with DVL3 in puncta (Fig. 6Db). Similar behavior was observed for DVL1 and DVL2 puncta (Appendix Fig. S3D). This demonstrates that TTLL11 can be actively recruited to the sites where phase separation of DVL3 takes place and as such can regulate it in situ. Altogether, this data shows that polyglutamylated DVL3 has a lower tendency to phase-separate and consequently forms more dynamic biomolecular condensates.

Polyglutamylation regulates phosphorylation of DVL3

The best-known regulators of DVL3 phase separation are several kinases—most notably CK1ε (Bryja et al, 2007) and centrosomal/ciliary kinases NEK2 and TTBK2 (Cervenka et al, 2016; Hanakova et al, 2019). These kinases upon binding and phosphorylation of DVL3 dissolve its condensates, resulting in more even cytosolic

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

© The Author(s)



Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

© The Author(s)

Figure 7. Deglutamylation of DVL3 by CCP enzymes.

(**A**) Co-localization of overexpressed Flag-DVL3 with EYFP-tagged CCP1, CCP5 and CCP6 in RPE cells. CCPs are recruited from microtubule/even distribution into DVL3 puncta. DVL3 is visualized using primary anti-Flag and corresponding secondary Alexa Fluor 568 antibodies. Scale bar = 20 μm. (**B**) Deglutamylation and interaction of Flag-DVL3 by/with CCPs. DVL3 variants (DVL3m stands for E710A/D714A mutant) were co-expressed with EYFP-tagged CCP1, CCP5 and CCP6 in HEK293T cells, lysed, immunoprecipitated using anti-Flag antibody (DVL3) and anti-GFP (CCPs) and analyzed by western blotting (WB). Polyglutamylation was detected by modification-specific antibody PolyE, quantified using ImageJ software and normalized on immunoprecipitated DVL3 protein amount (Flag antibody) in three biological replicates. Statistics were calculated by a one-sample *t* test with theoretical mean = 1. $^{**}P = 0.002$; $^{****}P < 0.0001$, ns not significant (**C**) Cells were transfected as in (**B**) but in DVL1/DVL2/DVL3 TKO HEK293T cells and subcellular localization of DVL3 was analyzed by immunocytochemistry. Subcellular localization of DVL3 was assessed for at least 200 cells in ($n \geq 3$) biological replicates. Mean cell counts and SD are indicated. Statistical significance in comparison to the condition without CCP for each DVL3 variant is shown. $^{**}P < 0.01$; $^{***}P < 0.001$, ns not significant. For statistical analysis, see "Methods" and source data files. Experiments from Figs. 6B,E and 7C were performed simultaneously, and control conditions are identical. Source data are available online for this figure.

distribution of DVL3 (Harnos et al, 2019). Interestingly, our interactome analysis (Fig. 4B,C) suggested increased interaction of CK1ε and CK1δ, a kinase closely related to and functionally redundant with CK1ε, with polyglutamylated DVL3. We further validated this observation by an independent approach—co-immunoprecipitation of endogenous CK1ε with overexpressed DVL3 in HEK293 cells was significantly promoted by co-expression with TTLL11 (Appendix Fig. S4A).

The obvious question then was "Are the polyglutamylation effects reflecting only the higher affinity of polyglutamylated DVL3 for CK1s or does polyglutamylation affect phase separation independently?". We have addressed the possible crosstalk between polyglutamylation and phosphorylation by the analysis of subcellular localization of DVL3 WT, polyglutamylated DVL3-12E, and glutamylation-resistant DVL3 E710A/D714A in the presence of exogenous CK1ε and in the presence of PF670462, the inhibitor of CK1δ and CK1ε. These experiments have been performed again in both *DVL1/DVL2/DVL3* TKO (Fig. 6E) and WT HEK293 cells (Appendix Fig. S4F). In both cases, co-expression of CK1ε strongly promoted uniform localization of all DVL3 variants, despite being less efficient in the case of DVL3 E710A/D714A (Fig. 6E). On the other side, CK1 inhibition, despite strongly promoting punctate distribution of DVL3, was not able to fully revert the even localization of DVL3-12E caused by polyglutamylation (Fig. 6E, bar 6). CK1 inhibition was less efficient when DVL3-12E was co-expressed with TTLL11 (see Appendix Fig. S4F). We conclude that both phosphorylation and polyglutamylation affect DVL3 phase separation independently, but their combination can be additive in vivo.

The observations of the additive effect of DVL3 C-terminal polyglutamylation and CK1ε-mediated phosphorylation raised the possibility that these events are mechanistically connected. To address this, we purified ^{15}N-labeled DVL3 C-terminal peptides corresponding to aa 693–716 (C24) and aa 693–716-12E (C24-12E) and analyzed the CK1ε phosphorylation kinetics of individual sites in vitro by real-time NMR (Fig. 6F,G; Appendix Fig. S4B–E). DVL3 polyglutamylation modulates CK1ε-mediated phosphorylation pattern in a qualitative and quantitative manner. In common, both peptides get initially phosphorylated at S709, but only the polyglutamylated variant gets fully phosphorylated at this site due to the much faster kinetics under identical experimental conditions. The rest three sites resident at the C-terminal tail, T695, S697, and S700, are phosphorylated only in the polyglutamylated form of the peptide, although partially. The in vitro data provide evidence that polyglutamylation enhances CK1ε activity and significantly affects the phosphorylation dynamics of the DVL3 C-terminus. These two modifications can thus regulate each other and control DVL functions connected with the phosphorylation of the C-terminus—including regulation of open and closed conformation (Harnos et al, 2019), participation in the noncanonical ROR2 signaling (Witte et al, 2010) or release from the centrosome during centriole separation (Cervenka et al, 2016).

DVL3 polyglutamylation and polyglutamylation-induced changes in phase separation are reversed by CCP6

Polyglutamylation of tubulin can be reversed by deglutamylating enzymes from the CCP protein family (Kimura et al, 2010; Rogowski et al, 2010; Tort et al, 2014). We hypothesized that if polyglutamylation of DVL3 has a biological significance, then a deglutamylation mechanism should exist to control this function. CCPs, which have previously been shown to deglutamylate not only tubulin but also non-tubulin proteins such as myosin light-chain kinase (MLCK), whose C-terminus consists of polyglutamate stretch (Rogowski et al, 2010), were the most likely candidates for such function. We thus tested CCP1, CCP5, and CCP6 as the best-characterized deglutamylating enzymes. CCP1 and CCP6 are deglutamylases shortening polyE chains (Rogowski et al, 2010), while CCP5 was proposed to remove γ-linked branching glutamylation points (Rogowski et al, 2010), albeit it can also remove α-linked glutamates at a slower rate (Berezniuk et al, 2013). As shown in Fig. 7A, all three CCP enzymes were localized to the cytoplasm when overexpressed, however at least a fraction was recruited into DVL3 puncta upon DVL3 co-expression. In particular, CCP6 showed almost complete co-localization with DVL3 in these puncta. This suggests that CCP proteins might be capable of interacting with DVL3.

To evaluate such interactions and the potential deglutamylation activity on DVL3 directly, we co-expressed CCP1, CCP5 and CCP6 together with DVL3 WT, the polyglutamylation-mimicking DVL3-12E, as well as glutamylation-resistant DVL3 E710A/D714A (Fig. 7B). First, this experiment showed that CCP6 binds DVL3 WT more than CCP1 and CCP5 (IP: FLAG, WB: GFP). Second, we detected a strong complex between DVL3-12E and CCP5 and CCP6 but not CCP1 (IP: GFP, WB: FLAG). On the other hand, DVL3 E710A/D714A showed weaker interaction with all CCPs. This suggests that the interaction of CCP5 and CCP6 is promoted by the C-terminal polyglutamate sequence of DVL3. Lastly and most importantly, CCP1 and especially CCP6 were able to efficiently reduce polyglutamylation of DVL3-12E (IP: Flag, WB: PolyE), which agrees with their known function of removing C-terminal glutamate residues of a linear peptide chain. By contrast, CCP5, known to preferentially act on polyglutamylation branching points (Berezniuk et al, 2013; Rogowski et al, 2010), was less efficient in deglutamylating DVL3. The role of CCP1 and CCP6 in DVL3 deglutamylation was further confirmed by showing that

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

© The Author(s)

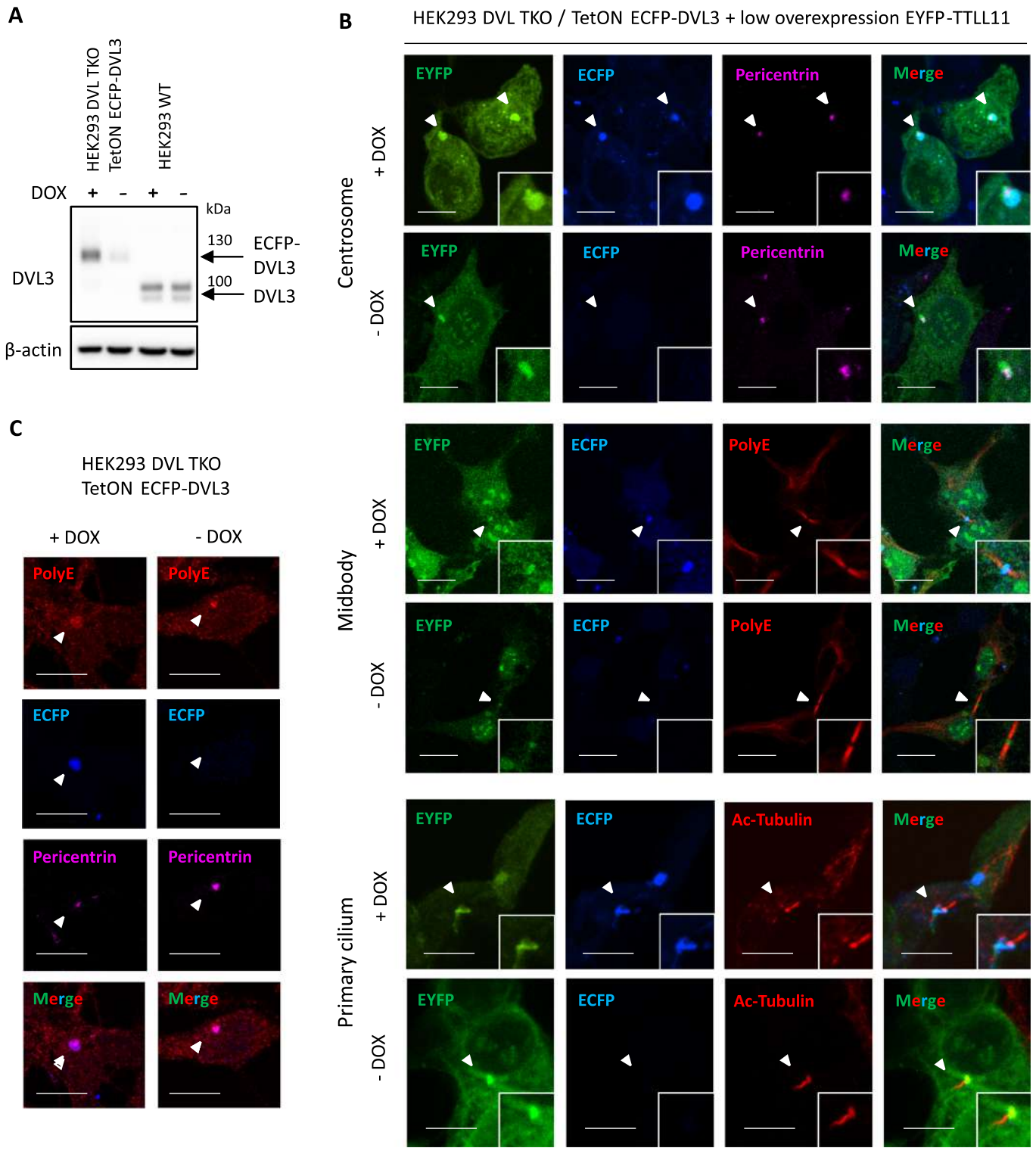


Figure 8. Subcellular co-localization of DVL3 with TTLL11.

(**A**) Stable cell line HEK293 DVL1/2/3 KO (triple KO-TKO) with inducible ECFP-DVL3 were transfected with low amount of EYFP-TTLL11 (20 ng) or control vector and doxycycline (DOX) induced near-endogenous level of ECFP-DVL3 was confirmed by WB analysis. (**B**) Co-localization of near-endogenous ECFP-DVL3 with TTLL11 in centrosome, midbody and primary cilium. Arrowheads point to the place of co-localization. (**C**) Polyglutamylation signal in in centrosome of cells with or without ECFP-DVL3. Arrowheads show centrosome. Micrographs shown are max-intensity projections of Z-stack taken to obtain whole microtubular compartments. Scale bar represents 10 μm. Source data are available online for this figure.

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

 © The Author(s)

both enzymes, and in particular CCP6, can efficiently reduce DVL3 polyglutamylation generated by TTLL11 (Appendix Fig. S5A). This effect is specific and depends on the CCP enzymatic activity, as inactive mutants of both CCPs did not reduce DVL3 polyglutamylation. In summary, we showed that CCP1 and CCP6 can remove polyglutamylation of DVL3, and from the tested CCPs, CCP6 is the most efficient DVL3 deglutamylase. We propose that this specificity is achieved by the combination of its capacity to bind polyglutamylated DVL3 stronger than CCP1, which is also deglutamylating DVL3, and to remove linear polyglutamate chains (in contrast to CCP5, which also binds polyglutamylated DVL3).

Knowing that CCPs can remove polyglutamylation of DVL3 we asked whether they also revert the effects polyglutamylation has on phase separation of DVL3. We analyzed the subcellular localization of WT DVL3 and polyglutamylation-mimicking DVL3-12E after co-expression with CCP1, CCP5, and CCP6 (Fig. 7C; Appendix Fig. S5B). Interestingly, while CCP1 and CCP5 did not significantly change the localization pattern of DVL3, CCP6 was able to promote its punctate phenotype, suggesting it promotes the formation of DVL3 protein condensates. This effect of CCP6 was even more pronounced in the case of DVL3-12E (Fig. 7C, compare bars 5–8). This demonstrates that CCP6 not only deglutamylates DVL3 but also prevents its polyglutamylation-induced cellular phenotype. We conclude that CCP6 acts as an enzyme that removes DVL3 polyglutamylation and counteracts the role of polyglutamylation/TTLLL11-mediated control of DVL3 phase separation.

DVL3 co-localizes with TTLL11 in microtubular structures

The top hits that are binding specifically to polyglutamylated DVL3 (Fig. 4; RAB11FIP, KATNAL2) are proteins associated with microtubules and microtubule-formed structures. Therefore, we tested if we could detect co-localization of TTLL11 with DVL3 in structures such as centrosome, primary cilia and midbody (Fig. 8). We took advantage of previously characterized HEK293 *DVL1/2/3* triple KO with the inducible expression of ECFP-DVL3 (Cervenka et al, 2016). We optimized doxycycline induction to obtain the near-endogenous level of ECFP-DVL3 expression (Fig. 8A). Since we could find no antibody capably to specifically detect endogenous TTLL11, we co-transfected low amounts of EYFP-TTLL11 (Fig. 8B). Interestingly, we observed co-localization of both proteins in droplet-like structures that were associated with centrosome, basal body of primary cilium, or the middle part of midbody in case of dividing cells. We could not detect co-localization in mitotic spindle. The polyglutamylation detected by polyE antibody was too strong in case of EYFP-TTLL11 expression, but we could detect PolyE signal overlapping with DVL3 droplet surrounding the centrosome without the need to overexpress TTLL11 (Fig. 8C). We conclude that TTLL11 and DVL3 co-localize in several microtubule-based compartments and propose that this is the place where DVL3 polyglutamylation takes place.

Discussion

Novel type of PTM

Here we discovered that a key protein of the Wnt signaling pathway, DVL3, is post-translationally polyglutamylated by TTLL11 through a previously unknown and unique mechanism, in which a polyglutamate chain is generated at the α-COOH of the terminal amino acid of the substrate protein. The direct elongation of primary peptide chains by adding a variable number of glutamates substantially increases the breadth of polyglutamylation as a PTM. Further, many more proteins than previously expected can be polyglutamylated, because, as we demonstrated, it can take place at multiple C-terminal amino acids including M, A, E, and S. Moreover, TTLL11 could tolerate big changes at the DVL3 C-terminus—including deletion of C-terminal regions or C-terminal elongation by a stretch of 4 As. We propose that once TTLL11 interacts with its substrate it can behave as a rather promiscuous enzyme. Only one sequence modification completely prevented polyglutamylation in our hands: the mutation of two acidic residues, E710 and D714 to A, which made the DVL3 C-terminus very hydrophobic. It thus appears that the charge/hydrophilicity of amino acid residues near the C-terminal modification site, rather than the sequence motif or identity of C-terminal amino acid, is important for TTLL11-mediated polyglutamylation. This is in agreement with structural analyses of the catalytic site of TTLL enzymes (Natarajan et al, 2017) and their recognition domain (Garnham et al, 2015) which are both positively charged. Our discovery opens the exciting possibility that many other proteins get polyglutamylated by this mechanism. This is supported by the observation that overexpression of TTLL11 ((van Dijk et al, 2007), this study) resulted in multiple polyE-positive bands on the WB of whole-cell lysates. It remains to be identified if/what are the other substrates of TTLL11 in addition to DVL3, and to what extent can other TTLLs elongate α-carboxyl of other amino acids than glutamate and what are their substrate proteins.

Novel biochemical/biophysical processes regulated by polyglutamylation

Given that the best-studied substrate of TTLL polyglutamylases is tubulin (Alexander et al, 1991; Edde et al, 1990; Redeker et al, 1991), initial insights into the function of polyglutamylation come from studies showing the role of this PTM in the modulation of interactions between microtubules and their interacting proteins, such as the microtubule severing proteins katanin and spastin (Lacroix et al, 2010; Shin et al, 2019; Valenstein and Roll-Mecak, 2016), or molecular motors (Sirajuddin et al, 2014). The identification of DVL3 as a polyglutamylation substrate allowed us to discover previously unknown, distinct functions of polyglutamylation, such as the control of protein phosphorylation and phase separation.

Our demonstration that polyglutamylation of DVL3 reduces its capacity to form biomolecular condensates phenocopies the effects of CK1ε—the established master regulator of DVL biology that controls its role in the Wnt signaling pathways (Bryja et al, 2007; Cong et al, 2004; Peters et al, 1999; Schwarz-Romond et al, 2005). These two mechanisms—polyglutamylation and phosphorylation of DVL3—probably act in parallel and are mutually supportive as CK1ε binds stronger to polyglutamylated DVL3 and polyglutamylated C-terminus of DVL3 gets phosphorylated by CK1ε much more efficiently. Both phosphorylation and polyglutamylation are adding a negative charge to the modification site. As charge distribution in IDPs significantly affects protein phase separation (Das and Pappu, 2013), disbalancing the amount and distribution of charged residues is likely the mechanism of polyglutamylation

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

© The Author(s)

and phosphorylation-mediated reduction in DVL3 phase separation capacity. Given the biochemical nature of polyglutamylation that allows for gradual charge and signal modulation—a regulatory principle so far only demonstrated for microtubule severing (Lacroix et al, 2010; Shin et al, 2019; Valenstein and Roll-Mecak, 2016)—we propose that polyglutamylation could play a unique role in fine-tuning of multiple phase separation processes where it can positively synergize with phosphorylation.

Phase separation of DVL—physiological relevance and role in Wnt signaling

DVL clearly phase separates, but the physiological significance of this process is still unclear. In the overexpression setup, several key determinants that promote (DVL DIX domain, DVL DEP domain, ubiquitination by WWP2, but also several regions in the IDRs (Axelrod et al, 1998; Gammons et al, 2016; Kang et al, 2022; Schwarz-Romond et al, 2005; Vamadevan et al, 2022)) or reduce (phosphorylation by CK1 or NEK2, (Bernatik et al, 2014; Cervenka et al, 2016; Mund et al, 2015)) DVL phase separation has been identified. The fact that these determinants and stimuli do not directly correlate with the activity in the Wnt/β-catenin pathway suggests that the phase separation of DVL and its signaling activity in the Wnt/β-catenin pathway cannot be directly connected.

Till now several studies proved the existence of DVL protein condensates on the endogenous level. Many of them observed DVL condensates by immunostaining during the development of various organisms (Axelrod et al, 1998; Hawkins et al, 2005; Henson et al, 2021; Torres and Nelson, 2000; Yanagawa et al, 1995) and several studies observed that the distribution of DVL droplets was polarized (Miller et al, 1999; Swartz et al, 2021a; Torres and Nelson, 2000). Endogenous tagging of DVL and subsequent imaging of plasma membrane by TIRF microscopy suggested formation of DVL oligomers containing only several DVL molecules after pathway activation (Kan et al, 2020; Ma et al, 2020). Different group used super-resolution microscopy and observed small cytosolic condensates, that increased in amount after canonical Wnt stimulation (Kang et al, 2022). This suggests that a signaling complex at the plasma membrane, often termed as signalosome, which forms after Wnt stimulation involves DVL oligomerization, but is unclear if it involves phase separation. In noncanonical Wnt pathway, immuno-TEM and super-resolution imaging in sea urchin embryos observed that endogenous DVL forms dynamic droplets in cytosol, near the plasma membrane in the vegetal cortex as a part of cell polarization and embryo axis-defying event (Henson et al, 2021; Swartz et al, 2021a) Moreover, imaging in fly wing proposed dissolution of ectopic DVL condensates as part of PCP (Axelrod et al, 1998). Finally, immunocytochemical staining of endogenous DVL in model cell lines (Cervenka et al, 2016; Xie et al, 2018), C-terminal tagging of endogenous DVL2 (Schubert et al, 2022) as well as near-endogenous expression of DVL3-ECFP (Fig. 8) showed a cytoplasmic droplet that was associated with centrosome. Therefore, there could be different types endogenous DVL condensates, whose features need to be addressed in future. However, we propose that polyglutamylation affects the centrosome and microtubule compartment-associated DVL, where it co-localizes with TTLL11.

Polyglutamylated DVL and microtubular compartments

DVL has been previously identified to be associated with several microtubular compartments such centrosome, basal body, cilium and midbody. One of the key functions of DVL is its role in the basal body and centrosome, summarized in (Bryja et al, 2017), where TTLL11 (van Dijk et al, 2007), as well as CCPs, including CCP6 (Rodríguez de la Vega Otazo et al, 2013) are also localized. Release of DVL3 from the centrosome is controlled by a centrosomal kinase NEK2 (Cervenka et al, 2016) that at the same time blocks phase separation of DVL3 (Hanakova et al, 2019). In the ciliated cells DVL, function in both, cell polarization and docking of basal body to the membrane (Park et al, 2008). DVL3 knockdown resulted in a significantly reduced ciliogenesis (Park et al, 2008; Zhang et al, 2023). Although we were unable to directly link DVL polyglutamylation and ciliogenesis it is intriguing that TTLL11 gets upregulated during primary cilium formation (Mathieu et al, 2021), RAB11FIP5 vesicles localize to the base of primary cilia (He et al, 2018) and KATNAL2 regulates ciliogenesis and brain development in *Xenopus* embryos (Ververis et al, 2016). Interestingly, RAB11FIP5 was shown to be necessary for the ciliary import of polyglutamylase TTLL5 (He et al, 2018). Given the fact that centrosome and pericentriolar matrix (PCM) is perceived as the phase-separated organelle (Zwicker et al, 2014), it will be exciting to explore to which extent polyglutamylation of non-tubulin proteins could, together with centrosomal kinases, control the dynamics of PCM clients via regulation of their phase separation propensity (Woodruff et al, 2017).

Polyglutamylation in the noncanonical Wnt signaling

It has been shown earlier that decreased phase separation correlated with the "open" conformation of DVL3 (Harnos et al, 2019), which happened when binding of the DVL3 C-terminus to its own PDZ domain was prevented. Intriguingly, the binding of the DVL3 C-terminus to its PDZ domain is mediated by α-carboxyl of M716 (Harnos et al, 2019; Lee et al, 2015). Interestingly, it was shown in *Xenopus* embryos that the addition of the C-terminal tag to DVL, which is expected to lead to the open conformation, promoted the activity of DVL in Wnt/PCP but not Wnt/β-catenin (Qi et al, 2017). This is in perfect agreement with our observations and suggests that TTLL enzymes mediating physiological polyglutamylation can function as important regulators of CE during embryonal development. Peak expression of xTTLL11 at the onset of gastrulation as well as known functions of several proteins that specifically interact with polyglutamylated DVL3 support this possibility. In addition to CK1ε with the known role in the Wnt/PCP pathway (Bryja et al, 2008; McKay et al, 2001; Peters et al, 1999), we also showed that KATNAL2 and RAB11FIP5, affected CE in *Danio rerio* development. KATNAL2 has a role in microtubule severing, and its knockdown in *Xenopus* embryos led to defects in CE including blastopore and neural tube closure (Willsey et al, 2018). Moreover, the same study showed that KATNAL2 is expressed in neurogenic tissues and subcellularly localized to the basal body, ciliary axoneme, centriole, and mitotic spindle. RAB11FIP5, a RAB11 effector involved in apical endosome recycling pathway phenocopied Wnt11 in *Xenopus*, whereas its loss resulted in CE defects (Wallkamm et al, 2016). This indirect evidence suggests that TTLL-mediated polyglutamylation can integrate multiple

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

 © The Author(s)

stimuli required for proper onset of activation of Wnt/PCP during gastrulation.

Conclusions/limitations

In summary, we here discovered polyglutamylation via α-carboxyl of amino acid other than glutamate, which can effectively lead to the C-terminal linear elongation of primary peptide chains of a substrate. The minimal sequence requirements of the modification site make C-terminal polyglutamylation a potentially versatile PTM with a broad range of substrates. Analyzing C-terminal polyglutamylation of the signaling protein DVL3, we discovered that polyglutamylation controls protein phosphorylation and formation of protein condensates. Polyglutamylation controls DVL3 activity in the Wnt/PCP pathway and early embryonal development. Our study keeps open several key questions that can become a basis for the follow-up work. First, TTLL11 is likely not the only enzyme that can catalyse DVL polyglutamylation. Downregulation of TTLL11 exhibited severe phenotype in *Xenopus*, while no phenotype was observed in *Danio rerio* (this study) or in mouse (see https://www.mousephenotype.org/data/genes/MGI:1921660; (Dickinson et al, 2016). The identification of other responsible TTLL enzymes can be challenging due to redundancy among individual TTLL family members that has been described earlier in other contexts, for example, for TTLL4 and TTLL6 (Casanova et al, 2015). Our study also does not provide a clear functional analysis in vivo using the rescue experiments. Generation of novel experimental models that will allow a detailed analysis of potential PCP defects in the best-defined tissue contexts such as polarity in the skin. Finally, because of the low levels of both DVL3 and TTLL11, the follow-up studies need to define the role of extent and function of polyglutamylated DVL at a completely endogenous level.

Methods

Cell culture, recombinant Wnt treatment, and near-endogenous expression of ECFP-DVL3

HEK293T (ATCC-CRL-11268), hTERT RPE-1 (kind gift from E. Nigg, Biozentrum, University of Basel, Basel, Switzerland), HEK293T-REx DVL1-2-3$^{-/-}$ and HEK293T-REx DVL1-2-3$^{-/-}$/RNF43$^{-/-}$/ZNRF3$^{-/-}$ (Paclíková et al, 2017; Paclíková et al, 2021) were grown at 37 °C and 5% CO_2 in complete DMEM medium containing 10% fetal bovine serum (DMEM; Gibco #41966-029), 2mM L-glutamine (Life Technologies, #25030024), 50 U/ml penicillin, and 50U/ml streptomycin (Hyclone-Biotech, #SV30010). Cells were routinely checked for mycoplasma contamination. Cell passaging of adherent cell lines was performed using trypsin (Biosera). The cells were washed in phosphate-buffered saline (PBS) buffer prior to the passage. For protein purification, suspension culture of FreeStyle™ 293-F Cells (Thermo Fisher Scientific) was cultured in FreeStyle™ 293 expression medium (Thermo Fisher Scientific) at 37 °C and 5% CO_2 while shaking at 120 rpm in Erlenmeyer cell culture flasks (Corning). For peptide purification, see "Peptide cloning, expression, and purification".

Treatment by recombinant Wnt3a and Wnt5a was performed as follows: HEK293T cells were treated with 1 μM Porcupine inhibitor LGK974 (Stem RD, #974-02) at the time of seeding and was maintained in the media whole time. Cells were transfected by corresponding vectors and 24 h post-transfection medium was supplemented with 100 ng/ml Wnt3a or Wnt5a (#1324-WN and #645-WN; R&D Systems) and cells were harvested after 3-h treatment.

For near-endogenous expression of ECFP-DVL3, stable cell line HEK293 DVL1/2/3 KO (triple KO–TKO) with inducible ECFP-DVL3 (Harnos et al, 2019) was treated with 0.5 μg/ml of doxycycline for 24 h. Primary cilia formation was achieved by changing cell media for a serum-free variant for 24 h. The samples for WB analysis of ECFP-DVL3 amount were prepared in the same way as the samples for immunostaining until the fixation step.

Transient transfection

List of plasmids used in this study is available in Dataset EV2. Adherent cells were seeded with density of 30,000 cells per cm^2 if not indicated otherwise (in case of immunocytochemical methods the density was 10,000 cells per cm^2). They were incubated for 24 h and transfected using Polyetylenimin (PEI) with stock concentration of 1 mg/ml and pH 7.4. The amount of PEI and DNA as well as the incubation time of cells with transfection mixture differed according to experiment Dataset EV3. DNA and PEI were separately mixed with serum-free DMEM, incubated for 20 min at room temperature (RT) and subsequently mixed. Each mixing step was followed by vortexing and centrifugation. The final mix was incubated for 20 min at RT and added to the cells. A medium was replaced with the fresh complete DMEM after 4-h-long incubation with transfection mixture. The cells were harvested or fixed 24 h after transfection. Suspension FreeStyle™ 293-F cells were transfected in 200 ml culture with density of 4×10^6 cells per 1 ml. 400 μg DNA was diluted in 7 ml of PBS, while 1.2 ml of PEI (1 mg/ml) was diluted in 3 ml of PBS. Both mixtures were incubated for 20 min and subsequently mixed and incubated for another 20 min before adding to cell culture. In total, 200 ml of fresh media was added to the culture 4 h post-transfection.

Dual-luciferase assay for analysis of Wnt/β-catenin signaling

Dual-Luciferase® Reporter Assay System (Promega) was performed by two reporter vectors. The test reporter vector (Super8X TopFlash) contained a gene for firefly luciferase controlled by the promoter containing TCF4-binding repeats. TCL/LEF together with β-catenin are co-activating subsequent transcription. The control reporter (pRLtkLuc) contained the gene for Renilla luciferase with the constitutively active promoter. Therefore, an activity of control reporter was used to normalize data for cell viability and transfection variability. The HEK293T cells were seeded on a 24-well plate and transfected according to Dataset EV3. For rescue experiment (Fig. EV4C), cells were treated with 1 μM Porcupine inhibitor LGK974 (Stem RD, #974-02) at the time of seeding. For Wnt stimulation, 200 μl of Wnt3a conditioned medium per condition was used in total 500 μl of medium per well for 14 h (as described in (Paclíková et al, 2021)). The cells were lysed after 24 h incubation (or 14 h in case of rescue experiment Fig. EV4C) and processed following the manufacturer's instructions. Luminescence was measured on MLX luminometer (Dynex Technologies).

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

© The Author(s)

Site-directed mutagenesis and cloning

The Twin-Strep-Flag-HALO-DVL3 wt plasmid was made by Gateway Technology (via LR recombination according to the manufacturer's instructions; Thermo Fisher Scientific, #11791020) from the donor plasmid pDONR221 DVL3 (DNASU, #FLH178665.01x) and the destination plasmid pDEST-Twin-Strep-Flag-HALO (kind gift from Cyril Bařinka). The Twin-Strep-Flag-ECFP-DVL3 wt was made as follows: pdECFP-DVL3 (Harnos et al, 2019) vector was used as template for PCR product that was inserted into pDONR221, resulting in pDONR221-ECFP-DVL3 vector that was combined with pDEST-Twin-Strep-Flag-HALO via Gateway Technology (Thermo Fisher Scientific #11791020). HALO tag was subsequently removed by site-directed mutagenesis. Mutagenesis of pcDNA3.1-Flag-hDVl3, Twin-strep-Flag-HALO-DVL3 or pCS2+xDvl3 vector was performed using QuikChange II XL Site-Directed Mutagenesis Kit (Agilent Technologies) following a manufacturer's instructions. Some vectors were made by serial mutagenesis as shown in Dataset EV3. The PCR was performed using gradient thermal cycler (Bio-Gener Technology). For a bacterial transformation, One Shot® TOP10 Chemically Competent E. coli (Invitrogen) and Super Optimal Broth (SOC) medium were used (Thermo Fisher Scientific). The primers designed for mutagenesis reactions are listed in Dataset EV4. All mutations were verified using Sanger sequencing. For peptide cloning, see "Peptide cloning, expression, and purification".

Quantitative PCR

RNA was purified by RNeasy Mini Kit (Qiagen). cDNA was synthesized following the manufacturer's instructions using the RevertAid Reverse Transcriptase kit (Thermo Fisher Scientific). qRT-PCR was conducted on a Roche LightCycler with the following program: initial denaturation at 95 °C for 10 min, followed by 45 cycles of 95 °C for 10 s, 60 °C for 30 s, and 72 °C for 1 s. Results were calculated as the fold difference between the target gene and the reference gene using the relative quantification $2^{-\Delta Cq}$ method. Transcript levels were measured using the LightCycler® FastStart DNA Master SYBR Green I kit (Roche) according to the manufacturer's protocol. 40S ribosomal protein S13 (RPS13) served as reference gene and primer sequences and probes are detailed in Dataset EV6.

Co-immunoprecipitation

The HEK293T cells were grown at 10-cm dishes to 60–70% confluency. They were transfected with desired vectors and incubated for 24 h post-transfection. The following work was performed at 4 °C. Cells were lysed for 20 min in 1 ml of lysis buffer (50 mM Tris, pH = 7.4, 150 mM NaCl, 1 mM EDTA, 0.5% NP40) supplemented by 1 mM DTT (Sigma-Aldrich) and 1× protease inhibitors (Roche Applied Science). Crude cell lysate was cleared by centrifugation for 15 min at 16,000 × *g*. The 80 µl of total cell lysate (TCL) was mixed with 20 µl of 5× Laemmli buffer. The remaining supernatant was divided into samples and each sample was incubated overnight with 1 µl of the corresponding antibody. G-protein coupled sepharose beads (GE Healthcare) were pre-washed in lysis buffer and blocked overnight in the lysis buffer with 1% BSA. The next day, lysates were combined with 15 µl of the pre-washed G-protein coupled sepharose beads. The mixture was incubated on carousel for 4 h. After the incubation, the beads were washed 5× in lysis buffer and proteins were eluted by the addition of Laemmli buffer to total volume of 80 µl. Used antibodies are listed in Dataset EV5.

Western blotting

For WB, the samples were incubated at 95 °C for 5 min and subjected to SDS-PAGE, then electrotransferred onto Hybond-P membrane (GE Healthcare), immunodetected using appropriate primary and secondary antibodies (conjugated with HRP) and visualized by ECL (Millipore) and documented using FusionSL system (Vilber-Lourmat). WB signal intensities were quantified using ImageJ. The area of the peak intensity for the protein of interest was divided by corresponding values of peaks intensity obtained for control protein. Used antibodies and corresponding dilutions are listed in the Dataset EV5.

Immunofluorescence

The HEK293T cells were seeded onto 13-mm coverslips, coated in 0.1% gelatin. Next day, the cells were transfected with vectors according to the design of experiments and if applicable treated by 1 µM PF670462 (DC chemicals, #DC2086), or the medium was changed for serum-free variant, and subsequently incubated for 24 h. Then, cells were washed in PBS, fixed in 4% paraformaldehyde (PFA, Millipore) in PBS for 1 h followed by three washes in PBS and finally blocked in PBTA (PBS, 5% donkey serum, 0.3% Triton X-100, 1% BSA). Samples were incubated overnight at 4 °C with primary antibodies diluted in PBTA. Following three washes in PBS, samples were incubated with corresponding Alexa Fluor secondary antibodies (Invitrogen, Abcam) for 1 h at RT, followed by 5 min incubation at RT with DAPI (Thermo Fisher) for nucleolar staining. Finally, samples were mounted in DAKO KO mounting solution (DAKO KO). The images were taken using a fluorescent (Olympus IX51) or confocal (Leica SP8) microscope using ×40 water or ×60 oil objectives. For protein subcellular localization experiments, at least 200 positive cells per experiment (n = 3) were analyzed and scored according to their phenotype into two categories (puncta/even). Used antibodies and corresponding dilutions are listed in Dataset EV5.

FRAP

RPE cells were seeded to µ-Slide 8 Well (Ibidi) chambers at a density of 10,000 cells per cm^2 24 h prior to transfection. Cells were transfected with corresponding plasmids of ECFP-tagged DVL3 construct (Datasets EV2 and EV3) and incubated for another 24 h. Live-cell imaging and photobleaching was performed on a Zeiss LSM 880 microscope equipped with a thermostat and CO_2 chamber using Plan-Apochromat ×63 objective. After transferring of cells into the microscope, they were left for 30 min to rest. Then, protein condensates of similar size were bleached using a 458-nm laser and images were taken every 5 s. Only condensates that stayed in the plane of focus were considered for analysis. Raw images were processed using ImageJ software using jay plugins (https://github.com/jayunruh/Jay_Plugins) from Jay Unruh at Stowers Institute for Medical Research in Kansas City, MO. Plugins used: create spectrum jru v1, combine all trajectories jru v1, normalize trajectories jru v1 (Min_Max normalization), batch FRAP fit jru v1 and average trajectories jru v1; all using default settings.

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

 © The Author(s)

LC-MS analysis of polyglutamylated DVL

Selected 1D gel bands were excised manually and after destaining and washing procedures each band was subjected to protein reduction (10 mM DTT in 25 mM $NaHCO_3$, 45 min, 56 °C, 750 rpm) and alkylation (55 mM IAA in 25 mM $NaHCO_3$; 30 min, laboratory temperature, 750 rpm). After further washing by 50% ACN/$NaHCO_3$ and pure ACN, the gel pieces were incubated with 125 ng trypsin (sequencing grade; Promega) in 50 mM $NaHCO_3$. The digestion was performed for 2 h at 40 °C on a Thermomixer (750 rpm; Eppendorf). Tryptic peptides were extracted into LC-MS vials by 2.5% formic acid (FA) in 50% ACN with the addition of polyethylene glycol (20,000; final concentration 0.001%, Stejskal et al, 2013) and concentrated in a SpeedVac concentrator (Thermo Fisher Scientific). Additional digestion of selected tryptic digests was conducted by adding 5.2 nmol of CNBr to a vacuum-dried digest reconstituted in 0.5 M HCl (16 h, 25 °C, 750 rpm).

LC-MS/MS analyses of peptide mixtures coming from in-gel digestions were done using RSLCnano system (Thermo Fisher Scientific) online connected to Impact II Qq-Time-Of-Flight mass spectrometer (Bruker, Bremen, Germany). Prior to LC separation, tryptic digests were online concentrated and desalted using trapping column (100 µm × 30 mm, 40 °C) filled with 3.5-µm X-Bridge BEH 130 C18 sorbent (Waters, Milford, MA, USA). After washing of trapping column with 0.1% FA, the peptides were eluted (flow 300 nl/min) from the trapping column onto an Acclaim Pepmap100 C18 column (3 µm particles, 75 µm × 500 mm, 40 °C; Thermo Fisher Scientific, Waltham, MA, USA) by 50 min nonlinear gradient program (1–56% of mobile phase B; mobile phase A: 0.1% FA in water; mobile phase B: 0.1% FA in 80% acetonitrile). The analytical column outlet was directly connected to the CaptiveSpray nanoBooster ion source (Bruker). NanoBooster was filled with acetonitrile and nanoBooster pressure was set to 0.2 Bar.

MS data were acquired in a data-dependent strategy with 3 s long cycle time. Mass range was set to 150–2200 *m/z* and precursors were selected from 300–2000 *m/z*. Acquisition speed of MS and MS/MS scans was 2 Hz and 4–16 Hz, respectively. Default CID collision energies and isolation widths with respect to precursor charge and *m/z* were used.

Under equivalent ionization and fragmentation conditions, peptide standards at a concentration 0.01 mg/ml (in 50% ACN, 0.1% FA) were directly injected at 3 µl/min flow rate to the ion source by using a syringe pump.

The processing of the mass spectrometric data including recalibration, compounds detection, charge deconvolution and further MS/MS data analysis was carried out in DataAnalysis software (4.2 SR1; Bruker). Mascot MS/MS ion searches (Matrixscience, London, UK; version 2.5.1) were done against in-house database containing expected protein sequences extended with polyE motifs added to the protein C-term or introduced into the sequence after sites tentatively forming side chains. cRAP contaminant database (downloaded from http://www.thegpm.org/crap/) was searched in advance to exclude contaminant spectra prior the main database search. Mass tolerance for peptides and MS/MS fragments were 10 ppm and 0.1 Da, respectively, with the option of one 13C atom to be present in the parent ion. Oxidation of methionine, carbamidomethylation (C), deamidation (N, Q) were set as optional modifications. In case of CNBr digests, methionine modification to homoserine or homoserine lactone was also allowed. All searches were done without enzyme specificity. Presence of modified C-terminal peptides was verified by checking the corresponding extracted ion chromatograms (± 0.015 *m/z*) and by manual verification of the MS/MS data obtained. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Perez-Riverol et al, 2022) partner repository with the dataset identifier PXD034237.

LC-MS analysis of protein complexes

LC-MS/MS analyses of all peptide mixtures were done using RSLCnano system connected to Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific). Prior to LC separation, tryptic digests were online concentrated and desalted using trapping column (100 µm × 30 mm) filled with 3.5-µm X-Bridge BEH 130 C18 sorbent (Waters). After washing of trapping column with 0.1% FA, the peptides were eluted (flow rate—300 nl/min) from the trapping column onto an analytical column (Acclaim Pepmap100 C18, 3 µm particles, 75 µm × 500 mm; Thermo Fisher Scientific) by 100 min nonlinear gradient program (1–56% of mobile phase B; mobile phase A: 0.1% FA in water; mobile phase B: 0.1% FA in 80% ACN). Equilibration of the trapping column and the column was done prior to sample injection to sample loop. The analytical column outlet was directly connected to the Digital PicoView 550 (New Objective) ion source with a sheath gas option and SilicaTip emitter (New Objective; FS360-20-15-N-20-C12) utilization. ABIRD (Active Background Ion Reduction Device, ESI Source Solutions) was installed (Paclíková et al, 2021).

MS data were acquired in a data-dependent strategy with cycle time for 3 s and with survey scan (350–2000 *m/z*). The resolution of the survey scan was 60,000 (200 *m/z*) with a target value of 4×10^5 ions and maximum injection time of 50 ms. HCD MS/MS (30% relative fragmentation energy, normal mass range) spectra were acquired with a target value of 5.0×10^4 and resolution of 30,000 (200 *m/z*). The maximum injection time for MS/MS was 50 ms. Dynamic exclusion was enabled for 60 s after one MS/MS spectra acquisition. The isolation window for MS/MS fragmentation was set to 1.6 *m/z* (Tanasa et al, 2023).

The analysis of the mass spectrometric RAW data files was carried out using the MaxQuant software (version 1.6.0.16) using default settings unless otherwise noted. MS/MS ion searches were done against modified cRAP database (based on http://www.thegpm.org/crap) containing protein contaminants like keratin, trypsin etc., and UniProtKB protein database for Homo Sapiens (https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/Eukaryota/UP000005640/UP000005640_9606.fasta.gz; downloaded 22.11.2017, version 2017/11, number of protein sequences: 21,009). Oxidation of methionine and proline, deamidation (N, Q), and acetylation (protein N-terminus) as optional modification, and trypsin/P enzyme with 2 allowed miscleavages were set. Peptides and proteins with FDR threshold <0.01 and proteins having at least one unique or razor peptide were considered only. Match between runs was set among all analyzed samples. Protein abundance was assessed using protein intensities calculated by MaxQuant (Tanasa et al, 2023). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Perez-Riverol et al, 2022) partner repository with the dataset identifier PXD033548.

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

© The Author(s)

Interactome analysis

Protein intensities reported in proteinGroups.txt file (output of MaxQuant) were further processed using the software container environment (https://github.com/OmicsWorkflows), version 3.7.2a. Processing workflow is available upon request. Briefly, it covered: (a) removal of decoy hits and contaminant protein groups, (b) protein group intensities $\log_2$ transformation, (c) LoessF normalization, (d) imputation by the global minimum and (e) differential expression using LIMMA statistical test. The volcano plot was created in R version 3.6.1, using the log fold change and adjusted *P* value reported by LIMMA. Proteins passing the threshold of log fold change >1 and adjusted *P* value <0.05 were further processed by REPRINT with integrated CRAPOME and SAINTexpress tools. For the background list of contaminants, CC66 was used; SAINTexpress was run with the default settings only against the CRAPOME control. REPRINT output was further processed by ProHits-Viz web interface producing Dot Plots and Correlation heatmaps using default settings. Scripts to reproduce analyses are available at Zenodo https://doi.org/10.5281/zenodo.11197196.

Recombinant protein purification

Full-length DVL3 wt, DVL3-E, DVL3 DCA (E710A/D714A) and TTLL11 (212-800) were produced as described in (31) with minimal changes: Twin-Strep-Flag-HALO N-terminally tagged DVL3 or TTLL11 were expressed in HEK293 cells using transient transfection. Cells were harvested 48 h post-transfection, resuspended in a lysis buffer (50 mM Tris pH 8, 150 mM NaCl, 10 mM KCl, 10% glycerol) with cocktail of protease inhibitors (#11836145001, Roche) and 0.2% NP40 (#74385, Sigma). The mixture was incubated for 20 min on ice and cell lysis was enhanced by dounce. Cell lysate was cleared by centrifugation at 100,000× *g* for 45 min at 4 °C and supernatant was loaded on Strep-Tactin Superflow high-capacity column (#2-1237-001, IBA) equilibrated in the lysis buffer. The column was washed in lysis buffer, and the protein was eluted using lysis buffer supplemented with 3 mM desthiobiotin. Eluted proteins were concentrated to 0.5 mg/ml using protein concentrators (#88516; Thermo Fisher Scientific), flash frozen in liquid nitrogen, and aliquots were stored at −80 °C. Tubulin and assembled MT were prepared as previously described (Skultetyova et al, 2017).

Peptide cloning, expression, and purification

pET vectors expressing C-terminal peptides of Dvl3 (C24 or C24-12E) were constructed using annealed oligo cloning. Complementary oligonucleotides were designed to create a double-stranded fragment with 5′ overhangs compatible with NcoI and KpnI restriction enzymes sites. Overlapping oligonucleotides were annealed, phosphorylated at 5′-termini by T4 Polynucleotide Kinase (M0201, New England Biolabs) and ligated using T4 DNA Ligase (M0202, New England Biolabs) to the pET vector digested with the same restriction enzymes. Finally, the expression constructs of the following form were obtained: His_6–ZZ tag–TEV cleavage site–peptide sequence. Peptides were expressed in *E. coli* BL21(DE3) strain (C2527I, New England Biolabs) and grown in M9 medium for ^{15}N labeling. The expression was induced with 0.5 mM IPTG (Alchimica). After cell harvest and lysis, the supernatant was loaded on an IMAC column (HisTrap HP 5 ml, GE Healthcare) in a buffer containing 25 mM Tris pH 8.0, 500 mM NaCl, 10 mM imidazole, 5% glycerol and eluted in the same buffer with 500 mM imidazole. The eluant was treated with TEV protease (produced in-house) for 1 h at 25 °C to cleave the tag and dialyzed for 16 h at 4 °C to remove the imidazole. The next day, the sample was boiled, snap cooled, and spined at 4000 × *g* for 20 min at 25 °C before loading to IMAC for tag removal. The flow-through containing the peptide was concentrated by centrifugation (Vivaspin 20, 3000 MWCO PES, Sartorius) and subjected to SEC (Superdex 75 10/300 GL, GE Healthcare) in NMR buffer (50 mM phosphate buffer pH 6.5, 50 mM KCl). Peptides were detected in SEC using low UV wavelength (214 nm). The CK1ε kinase core (1–301) was expressed in LB medium and purified as described before for CK1δ kinase core (Long et al, 2012).

Nuclear magnetic resonance spectroscopy

All experiments were recorded at 25 °C. Chemical shift assignment of the ^{15}N-labeled peptides was performed using a combination of ^{15}N-NOESY-HSQC and ^{15}N-TOCSY-HSQC experiments recorded at a 950 MHz Bruker Avance NEO spectrometer. Phosphorylation kinetics were measured using a series of ^{1}H-^{15}N HSQC experiments at a 700 MHz Bruker Avance NEO spectrometer. The general sample composition of a real-time phosphorylation measurement was as follows: ^{15}N-labeled peptide (50 µM), ATP (1 mM), $MgCl_2$ (10 mM), d_6-EDTA (1 mM), D_2O for lock (10%), CK1ε kinase core (1 µM; kinase:peptide ratio of 1:50) in NMR buffer up to a final volume of 200 µl. Kinase was added last in the NMR tube and series of ^{1}H-^{15}N HSQC spectra were recorded every 30 min over a time course of 16 h.

Nuclear magnetic resonance kinetic analysis

NMR spectra were analyzed using Sparky 3.115. The fraction of phosphorylation for a given amino acid at a given time point was calculated by dividing the peak intensity of the phosphorylated form by the sum of peak intensities of phosphorylated and non-phosphorylated forms ($I_{phospho}/(I_{phospho} + I_{nonphospho})$). The experimental data were fitted to a mono-exponential function f(x) = A*(1-exp(-k*x)) using Gnuplot 5.2.4.

In vitro polyglutamylation

DVL3 was mixed with TTLL11 (212-800) in BRB30 Buffer (30 mM PIPES, 8 mM MgCl2, 1 mM EGTA, pH 6.8) supplemented by 4 mM ATP, 4 mM Glu, 2 mM TCEP, 2×PI, 1 mM EDTA, TEV protease 0.03 mg/ml. Final protein concentration was 0.75 µM DVL, 0.75 µM MTs and 0.15 µM TTLL11. The mixture was either mixed with the sampling buffer or incubated 60 min at 37 °C and subsequently mixed with the sampling buffer. The proteins were separated using SDS-PAGE and visualized by CBB staining or immunoblotting. MS analysis was performed on the samples from excised bands of interest from CBB-stained gel.

Statistics

In graphs, numerical data are shown as mean ± SD as depicted. One-way ANOVA with Dunnett' multiple comparisons test was used for dual-luciferase assay in Fig. EV4C. For comparisons of fold

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

 © The Author(s)

change, one-sample t test with theoretical mean = 1 was used as samples were normalized to control (= 1) which as a result had zero variability ($^{*}P \leq 0.05$; $^{**}P \leq 0.01$; $^{***}P \leq 0.001$; ns, not significant, $P > 0.05$). Indicated statistical tests were calculated by GraphPad Prism 9 (GraphPad Software Inc.).

For DVL3 subcellular localization experiments, the data on frequencies puncta vs. even were analyzed using linear mixed model fit by REML and Type III ANOVA with Satterthwaite's method followed by multiple comparisons of means (Tukey contrast) in R software version 3.6.1 using lme4 and multcomp packages. Other method-specific statistical operations are explained in the corresponding method section. Scripts to reproduce analyses are available at Zenodo https://doi.org/10.5281/zenodo.11197197.

Animal models

The work with *Xenopus laevis* and *Zebrafish* was carried out according to the Czech animal use and care law and approved by the local authorities and committees (Animal Care and Housing Approval: 45055/2020-MZE-18134, issued by Ministry of Agriculture of the Czech Republic; and Animal Experiments Approval: CZ 62760214, issued by State Veterinary Administration/Section for South Moravian Region).

The adult *Xenopus laevis* frogs used for embryo generation are housed in the local husbandry facility at Masaryk University in a XenopLus system (catalog number RE18001301, Tecniplast, Italy), which is an advanced, fully automated system designed for amphibians, ensuring excellent animal care and housing conditions. The Xenopus laevis adults were purchased from the European Xenopus Research Center (ERXC) in 2021.

Zebrafish strain AB (ZDB-GENO-960809-7) was kept in the Techniplast ZebTec automatic recirculating system. They are fed twice a day with Artemia salina and Zebrafeed dry food by Sparos. The animals live under a 14/10-h day/night cycle and are kept in 28 °C water. After collecting the eggs and injections, the embryos were kept in E3 medium in an incubator at 28.5 °C until 14.5 hpf.

Only *Xenopus* and *Zebrafish* embryos at early developmental stages were studied, and no adult animals were involved in these experiments.

Xenopus embryos

The generation and cultivation of Xenopus embryos were performed following general protocols. In brief, testes from males under anesthesia (20% MS-222, Sigma-Aldrich, #A5040) were removed surgically from the body cavity and transferred to cold 1× Marc's Modified Ringers (MMR; 100 mM NaCl, 2 mM KCl, 1 mM $MgSO_4$, 2 mM $CaCl_2$, 5 mM HEPES, buffered to pH 7.4) supplemented with 50 µg/mL of gentamycin (Sigma-Aldrich, #G3632). To induce egg laying, fully mature Xenopus females were injected with 260 U of human chorionic gonadotropin (hCG; Merck, Ovitrelle 250 G) into the dorsal lymph sac about 12–16 h before use and kept overnight at 18 °C. For fertilization, eggs were squeezed from an induced female directly into a petri dish and mixed with a piece of testes in 0.1× MMR. After 20 min, embryos were dejellying using 3% cysteine solution (Sigma-Aldrich, C7880) buffered to pH 7.8 for circa 10 min until they contact each other. Washed embryos (5× in 0.1× MMR solution) were cultivated at 16–18 °C for subsequent microinjections of nucleic acids.

The RNA for microinjections was synthesized from the plasmid pCS2 myc-xDvl3 WT and its derivate 12E using the mMessage mMachine SP6 kit (Thermo Fisher, #AM1340). xTTLL11 ATG MO (taactttcttatgggcttttcccag) and xTTLL11 splicing MO for exon 3 (tacagaataatgagggccactca) were ordered from Gene tools (https://www.gene-tools.com/). For microinjections, embryos were cultivated in 3% Ficoll PM400 (Cytiva, #17030010) in 0.6× MMR and injected at the four-cell stage into the two dorsal blastomeres with 100 pg of myc-xDvl3 RNAs for the analysis of morphogenetic defects. Before gastrulation (stage 10.5), the embryos were cultivated in 0.1× MMR until they reached stage 26, fixed for 2 h in formaldehyde solution 4% buffered to pH 6.9 (Sigma-Aldrich, #100496), and analyzed for morphogenetic defects. Alternatively, five embryos (stage 26) were lysed before fixation in a 5× Laemmli loading buffer for the subsequent SDS- PAGE/WB analysis. As for the analysis of blastopore closure, the embryos were cultivated in 0.1× MMR enriched with 50 µg/mL of gentamycin (Sigma-Aldrich, #G3632). The number of embryos for each condition is indicated below the graphs (created by GraphPad software) and it is representative of 1–3 biological experiments. All embryos were staged after the normal table of Niewkoop and Faber (Nieuwkoop and Faber, 1994). The pictures of embryos were taken on plates with 2% agarose (Serva, #11404) using Leica S9i stereoscope and LAX software.

Zebrafish crispants

Oligonucleotide sequences for gRNA production were selected using the web tool https://chopchop.cbu.uib.no (Labun et al, 2019). Sequences of gRNA and oligos for sequencing are in the Dataset EV6. A protocol published by Shah et al (Shah et al, 2015) was utilized for gRNA production. A mixture consisting of rCas9NLS protein (0.66 mg/ml, Sigma-Aldrich CAS9PL), Phenol red (0.05%, Sigma-Aldrich, P0290), 120 mM KCl, and gRNA (200 ng/µl) was injected in a volume of 1 nl into the cells of one-cell stage embryos of the AB zebrafish strain. In control injections, an equivalent concentration of BSA was used in place of rCAS9 in an otherwise identical mixture. At 14.5 hpf, dechorionated embryos were fixed using 4% PFA for one hour, followed by three 20-minute washes in PBST (1% Tween in PBS). Subsequently, samples were permeabilized for 10 min in 1% Triton X-100 in PBS, washed for 20 min in PBST and blocked for 2 h in 3% BSA in PBST. Following blocking, the myoD1 antibody (Abcam ab209976, 1:400) was added and incubated overnight at 4 °C on a slow shaker. Then, samples were washed 3× for 20 min in PBST and incubated with a secondary antibody (α-Rabbit Alexa Fluor 488, 1:500) for 2 h at room temperature. After incubation, a final round of washing was performed. Finally, samples were dissected, mounted on slides, and imaged using a Leica SP8 confocal microscope.

Graphics

Graphics in Figs. 2B, 4A, 5A, 5E, 6F, EV4A, EV4, and synopsis image were created with BioRender.com

Data availability

The datasets and computer code produced in this study are available in the following databases: LC-MS analysis of DVL3 polyglutamylation data: PRIDE PXD034237; Polyglutamylated

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

© The Author(s)

DVL3 interactome data: PRIDE PXD033548; Scripts to reproduce analysis: Zenodo 11197197. All reagents used are mentioned in "Methods" or listed in supplementary material with corresponding source and/or RRID. Unique reagents generated in this study (plasmids, recombinant proteins) are available upon request.

The source data of this paper are collected in the following database record: biostudies:S-SCDT-10_1038-S44318-024-00254-7.

Expanded view data, supplementary information, appendices are available for this paper at https://doi.org/10.1038/s44318-024-00254-7.

Peer review information

A peer review file is available at https://doi.org/10.1038/s44318-024-00254-7

References

Alexander JE, Hunt DF, Lee MK, Shabanowitz J, Michel H, Berlin SC, MacDonald TL, Sundberg RJ, Rebhun LI, Frankfurter A (1991) Characterization of posttranslational modifications in neuron-specific class III beta-tubulin by mass spectrometry. Proc Natl Acad Sci USA 88:4685–4689

Axelrod JD, Miller JR, Shulman JM, Moon RT, Perrimon N (1998) Differential recruitment of Dishevelled provides signaling specificity in the planar cell polarity and Wingless signaling pathways. Genes Dev 12:2610–2622

Berezniuk I, Lyons PJ, Sironi JJ, Xiao H, Setou M, Angeletti RH, Ikegami K, Fricker LD (2013) Cytosolic carboxypeptidase 5 removes alpha- and gamma-linked glutamates from tubulin. J Biol Chem 288:30445–30453

Bernatik O, Ganji RS, Dijksterhuis JP, Konik P, Cervenka I, Polonio T, Krejci P, Schulte G, Bryja V (2011) Sequential activation and inactivation of Dishevelled in the Wnt/beta-catenin pathway by casein kinases. J Biol Chem 286:10396–10410

Bernatik O, Sedova K, Schille C, Ganji RS, Cervenka I, Trantirek L, Schambony A, Zdrahal Z, Bryja V (2014) Functional analysis of dishevelled-3 phosphorylation identifies distinct mechanisms driven by casein kinase 1ε and frizzled5. J Biol Chem 289:23520–23533

Bilic J, Huang YL, Davidson G, Zimmermann T, Cruciat CM, Bienz M, Niehrs C (2007) Wnt induces LRP6 signalosomes and promotes dishevelled-dependent LRP6 phosphorylation. Science 316:1619–1622

Boutros M, Paricio N, Strutt DI, Mlodzik M (1998) Dishevelled activates JNK and discriminates between JNK pathways in planar polarity and wingless signaling. Cell 94:109–118

Bryja V, Schambony A, Cajánek L, Dominguez I, Arenas E, Schulte G (2008) Beta-arrestin and casein kinase 1/2 define distinct branches of non-canonical WNT signalling pathways. EMBO Rep 9:1244–1250

Bryja V, Schulte G, Rawal N, Grahn A, Arenas E (2007) Wnt-5a induces Dishevelled phosphorylation and dopaminergic differentiation via a CK1-dependent mechanism. J Cell Sci 120:586–595

Bryja V, Červenka I, Čajánek L (2017) The connections of Wnt pathway components with cell cycle and centrosome: side effects or a hidden logic? Crit Rev Biochem Mol Biol 52:614–637

Butler MT, Wallingford JB (2017) Planar cell polarity in development and disease. Nat Rev Mol Cell Biol 18:375–388

Casanova M, de Monbrison F, van Dijk J, Janke C, Pagès M, Bastien P (2015) Characterisation of polyglutamylases in trypanosomatids. Int J Parasitol 45:121–132

Cervenka I, Valnohova J, Bernatik O, Harnos J, Radsetoulal M, Sedova K, Hanakova K, Potesil D, Sedlackova M, Salasova A et al (2016) Dishevelled is a NEK2 kinase substrate controlling dynamics of centrosomal linker proteins. Proc Natl Acad Sci USA 113:9304–9309

Choi H, Liu G, Mellacheruvu D, Tyers M, Gingras AC, Nesvizhskii AI (2012) Analyzing protein-protein interactions from affinity purification-mass spectrometry data with SAINT. Curr Protoc Bioinformatics Chapter 8:Unit8.15

Clevers H, Nusse R (2012) Wnt/β-catenin signaling and disease. Cell 149:1192–1205

Cong F, Schweizer L, Varmus H (2004) Casein kinase Iepsilon modulates the signaling specificities of Dishevelled. Mol Cell Biol 24:2000–2011

Das RK, Pappu RV (2013) Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proc Natl Acad Sci USA 110:13392–13397

de Groot RE, Ganji RS, Bernatik O, Lloyd-Lewis B, Seipel K, Šedová K, Zdráhal Z, Dhople VM, Dale TC, Korswagen HC et al (2014) Huwe1-mediated ubiquitylation of Dishevelled defines a negative feedback loop in the Wnt signaling pathway. Sci Signal 7:ra26

Devenport D (2014) The cell biology of planar cell polarity. J Cell Biol 207:171–179

Dickinson ME, Flenniken AM, Ji X, Teboul L, Wong MD, White JK, Meehan TF, Weninger WJ, Westerberg H, Adissu H et al (2016) High-throughput discovery of novel developmental phenotypes. Nature 537:508–514

Edde B, Rossier J, Le Caer JP, Desbruyeres E, Gros F, Denoulet P (1990) Posttranslational glutamylation of alpha-tubulin. Science 247:83–85

Fiedler M, Mendoza-Topaz C, Rutherford TJ, Mieszczanek J, Bienz M (2011) Dishevelled interacts with the DIX domain polymerization interface of Axin to interfere with its function in down-regulating beta-catenin. Proc Natl Acad Sci USA 108:1937–1942

Gammons MV, Renko M, Johnson CM, Rutherford TJ, Bienz M (2016) Wnt signalosome assembly by DEP domain swapping of Dishevelled. Mol Cell 64:92–104

Gao C, Chen YG (2010) Dishevelled: the hub of Wnt signaling. Cell Signal 22:717–727

Garnham CP, Vemu A, Wilson-Kubalek EM, Yu I, Szyk A, Lander GC, Milligan RA, Roll-Mecak A (2015) Multivalent microtubule recognition by tubulin tyrosine ligase-like family glutamylases. Cell 161:1112–1123

Hanakova K, Bernatik O, Kravec M, Micka M, Kumar J, Harnos J, Ovesna P, Paclikova P, Radsetoulal M, Potesil D et al (2019) Comparative phosphorylation map of Dishevelled 3 links phospho-signatures to biological outputs. Cell Commun Signal 17:170

Harnos J, Canizal MCA, Jurasek M, Kumar J, Holler C, Schambony A, Hanakova K, Bernatik O, Zdrahal Z, Gomoryova K et al (2019) Dishevelled-3 conformation dynamics analyzed by FRET-based biosensors reveals a key role of casein kinase 1. Nat Commun 10:1804

Harrison C, Shao H, Strutt H, Strutt D (2020) Molecular mechanisms mediating asymmetric subcellular localisation of the core planar polarity pathway proteins. Biochem Soc Trans 48:1297–1308

Hashimoto M, Shinohara K, Wang J, Ikeuchi S, Yoshiba S, Meno C, Nonaka S, Takada S, Hatta K, Wynshaw-Boris A et al (2010) Planar polarization of node cells determines the rotational axis of node cilia. Nat Cell Biol 12:170–176

Hawkins NC, Ellis GC, Bowerman B, Garriga G (2005) MOM-5 frizzled regulates the distribution of DSH-2 to control C. elegans asymmetric neuroblast divisions. Dev Biol 284:246–259

He K, Ma X, Xu T, Li Y, Hodge A, Zhang Q, Torline J, Huang Y, Zhao J, Ling K et al (2018) Axoneme polyglutamylation regulated by Joubert syndrome protein ARL13B controls ciliary targeting of signaling molecules. Nat Commun 9:3310

Henson JH, Samasa B, Shuster CB, Wikramanayake AH (2021) The nanoscale organization of the Wnt signaling integrator Dishevelled in the vegetal cortex domain of an egg and early embryo. PLoS ONE 16:e0248197

Hikasa H, Sokol SY (2013) Wnt signaling in vertebrate axis specification. Cold Spring Harb Perspect Biol 5:a007955

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

 © The Author(s)

Humphries AC, Mlodzik M (2018) From instruction to output: Wnt/PCP signaling in development and cancer. Curr Opin Cell Biol 51:110–116

Janke C, Magiera MM (2020) The tubulin code and its role in controlling microtubule properties and functions. Nat Rev Mol Cell Biol 21:307–326

Janke C, Rogowski K, Wloga D, Regnard C, Kajava AV, Strub JM, Temurak N, van Dijk J, Boucher D, van Dorsselaer A et al (2005) Tubulin polyglutamylase enzymes are members of the TTL domain protein family. Science 308:1758–1762

Kan W, Enos MD, Korkmazhan E, Muennich S, Chen DH, Gammons MV, Vasishtha M, Bienz M, Dunn AR, Skiniotis G et al (2020) Limited dishevelled/Axin oligomerization determines efficiency of Wnt/beta-catenin signal transduction. eLife 9:e55015

Kang K, Shi Q, Wang X, Chen YG (2022) Dishevelled phase separation promotes Wnt signalosome assembly and destruction complex disassembly. J Cell Biol 221:e202205069

Kikuchi K, Niikura Y, Kitagawa K, Kikuchi A (2010) Dishevelled, a Wnt signalling component, is involved in mitotic progression in cooperation with Plk1. EMBO J 29:3470–3483

Kimura Y, Kurabe N, Ikegami K, Tsutsumi K, Konishi Y, Kaplan OI, Kunitomo H, Iino Y, Blacque OE, Setou M (2010) Identification of tubulin deglutamylase among *Caenorhabditis elegans* and mammalian cytosolic carboxypeptidases (CCPs). J Biol Chem 285:22936–22941

Kishida M, Hino SI, Michiue T, Yamamoto H, Kishida S, Fukui A, Asashima M, Kikuchi A (2001) Synergistic activation of the Wnt signaling pathway by Dvl and casein kinase Iepsilon. J Biol Chem 276:33147–33155

Klingensmith J, Nusse R, Perrimon N (1994) The Drosophila segment polarity gene Dishevelled encodes a novel protein required for response to the wingless signal. Genes Dev 8:118–130

Koca Y, Collu GM, Mlodzik M (2022) Wnt-Frizzled planar cell polarity signaling in the regulation of cell motility. Curr Top Dev Biol 150:255–297

Labun K, Montague TG, Krause M, Torres Cleuren YN, Tjeldnes H, Valen E (2019) CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. Nucleic Acids Res 47:W171–W174

Lacroix B, van Dijk J, Gold ND, Guizetti J, Aldrian-Herrada G, Rogowski K, Gerlich DW, Janke C (2010) Tubulin polyglutamylation stimulates spastin-mediated microtubule severing. J Cell Biol 189:945–954

Lee HJ, Shi DL, Zheng JJ (2015) Conformational change of Dishevelled plays a key regulatory role in the Wnt signaling pathways. eLife 4:e08142

Long A, Zhao H, Huang X (2012) Structural basis for the interaction between casein kinase 1 delta and a potent and selective inhibitor. J Med Chem 55:956–960

Ma WZ, Chen MR, Kang H, Steinhart Z, Angers S, He X, Kirschner MW (2020) Single-molecule dynamics of Dishevelled at the plasma membrane and Wnt activation. Proc Natl Acad Sci USA 117:16690–16701

Mahalingan KK, Keith Keenan E, Strickland M, Li Y, Liu Y, Ball HL, Tanner ME, Tjandra N, Roll-Mecak A (2020) Structural basis for polyglutamate chain initiation and elongation by TTLL family enzymes. Nat Struct Mol Biol 27:802–813

Mathieu H, Patten SA, Aragon-Martin JA, Ocaka L, Simpson M, Child A, Moldovan F (2021) Genetic variant of TTLL11 gene and subsequent ciliary defects are associated with idiopathic scoliosis in a 5-generation UK family. Sci Rep 11:11026

McKay RM, Peters JM, Graff JM (2001) The casein kinase I family in Wnt signaling. Dev Biol 235:388–396

Mellacheruvu D, Wright Z, Couzens AL, Lambert JP, St-Denis NA, Li T, Miteva YV, Hauri S, Sardiu ME, Low TY et al (2013) The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. Nat Methods 10:730–736

Miller JR, Rowning BA, Larabell CA, Yang-Snyder JA, Bates RL, Moon RT (1999) Establishment of the dorsal-ventral axis in Xenopus embryos coincides with the dorsal enrichment of Dishevelled that is dependent on cortical rotation. J Cell Biol 146:427–437

Mund T, Graeb M, Mieszczanek J, Gammons M, Pelham HR, Bienz M (2015) Disinhibition of the HECT E3 ubiquitin ligase WWP2 by polymerized Dishevelled. Open Biol 5:150185

Natarajan K, Gadadhar S, Souphron J, Magiera MM, Janke C (2017) Molecular interactions between tubulin tails and glutamylases reveal determinants of glutamylation patterns. EMBO Rep 18:1013–1026

Nieuwkoop PD, Faber J (1994) Normal table of Xenopus laevis (Daudin): a systematical and chronological survey of the development from the fertilized egg till the end of metamorphosis. Routledge, New York

Paclíková P, Bernatík O, Radaszkiewicz TW, Bryja V (2017) The N-terminal part of the Dishevelled DEP domain is required for Wnt/β-catenin signaling in mammalian cells. Mol Cell Biol 37:e00145–17

Paclíková P, Radaszkiewicz TW, Potěšil D, Harnoš J, Zdráhal Z, Bryja V (2021) Roles of individual human Dishevelled paralogs in the Wnt signalling pathways. Cell Signal 85:110058

Park TJ, Mitchell BJ, Abitua PB, Kintner C, Wallingford JB (2008) Dishevelled controls apical docking and planar polarization of basal bodies in ciliated epithelial cells. Nat Genet 40:871–879

Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S, Kundu DJ, Prakash A, Frericks-Zipper A, Eisenacher M et al (2022) The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res 50:D543–D552

Peters JM, McKay RM, McKay JP, Graff JM (1999) Casein kinase I transduces Wnt signals. Nature 401:345–350

Qi J, Lee HJ, Saquet A, Cheng XN, Shao M, Zheng JJ, Shi DL (2017) Autoinhibition of Dishevelled protein regulated by its extreme C terminus plays a distinct role in Wnt/beta-catenin and Wnt/planar cell polarity (PCP) signaling pathways. J Biol Chem 292:5898–5908

Redeker V, Le Caer JP, Rossier J, Prome JC (1991) Structure of the polyglutamyl side chain posttranslationally added to alpha-tubulin. J Biol Chem 266:23461–23466

Regnard C, Desbruyères E, Huet JC, Beauvallet C, Pernollet JC, Eddé B (2000) Polyglutamylation of nucleosome assembly proteins. J Biol Chem 275:15969–15976

Rodríguez de la Vega Otazo M, Lorenzo J, Tort O, Avilés FX, Bautista JM (2013) Functional segregation and emerging role of cilia-related cytosolic carboxypeptidases (CCPs). FASEB J 27:424–431

Rogowski K, van Dijk J, Magiera MM, Bosc C, Deloulme JC, Bosson A, Peris L, Gold ND, Lacroix B, Bosch Grau M et al (2010) A family of protein-deglutamylating enzymes associated with neurodegeneration. Cell 143:564–578

Schubert A, Voloshanenko O, Ragaller F, Gmach P, Kranz D, Scheeder C, Miersch T, Schulz M, Trümper L, Binder C et al (2022) Superresolution microscopy localizes endogenous Dvl2 to Wnt signaling-responsive biomolecular condensates. Proc Natl Acad Sci USA 119:e2122476119

Schwarz-Romond T, Fiedler M, Shibata N, Butler PJ, Kikuchi A, Higuchi Y, Bienz M (2007a) The DIX domain of Dishevelled confers Wnt signaling by dynamic polymerization. Nat Struct Mol Biol 14:484–492

Schwarz-Romond T, Merrifield C, Nichols BJ, Bienz M (2005) The Wnt signalling effector Dishevelled forms dynamic protein assemblies rather than stable associations with cytoplasmic vesicles. J Cell Sci 118:5269–5277

Schwarz-Romond T, Metcalfe C, Bienz M (2007b) Dynamic recruitment of axin by Dishevelled protein assemblies. J Cell Sci 120:2402–2412

Sear RP (2007) Dishevelled: a protein that functions in living cells by phase separating. Soft Matter 3:680–684

Session AM, Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S, Fukui A, Hikosaka A, Suzuki A, Kondo M et al (2016) Genome evolution in the allotetraploid frog Xenopus laevis. Nature 538:336-343

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

© The Author(s)

Shah AN, Davey CF, Whitebirch AC, Miller AC, Moens CB (2015) Rapid reverse genetic screening using CRISPR in zebrafish. Nat Methods 12:535–540

Sharma M, Castro-Piedras I, Simmons GE, Pruitt K (2018) Dishevelled: a masterful conductor of complex Wnt signals. Cell Signal 47:52–64

Shin SC, Im SK, Jang EH, Jin KS, Hur EM, Kim EE (2019) Structural and molecular basis for katanin-mediated severing of glutamylated microtubules. Cell Rep 26:1357–1367.e1355

Sirajuddin M, Rice LM, Vale RD (2014) Regulation of microtubule motors by tubulin isotypes and post-translational modifications. Nat Cell Biol 16:335–344

Skultetyova L, Ustinova K, Kutil Z, Novakova Z, Pavlicek J, Mikesova J, Trapl D, Baranova P, Havlinova B, Hubalek M et al (2017) Human histone deacetylase 6 shows strong preference for tubulin dimers over assembled microtubules. Sci Rep 7:11547

Stejskal K, Potěšil D, Zdráhal Z (2013) Suppression of peptide sample losses in autosampler vials. J Proteome Res 12(6):3057–3062

Swartz SZ, Tan TH, Perillo M, Fakhri N, Wessel GM, Wikramanayake AH, Cheeseman IM (2021a) Polarized Dishevelled dissolution and reassembly drives embryonic axis specification in sea star oocytes. Curr Biol 31:5633–5641.e5634

Swartz SZ, Tan TH, Perillo M, Fakhri N, Wessel GM, Wikramanayake AH, Cheeseman IM (2021b) Polarized Dishevelled dissolution and reassembly drives embryonic axis specification in sea star oocytes. Curr Biol 31:5633–5641.e5634

Tamai K, Zeng X, Liu C, Zhang X, Harada Y, Chang Z, He X (2004) A mechanism for Wnt coreceptor activation. Mol Cell 13:149–156

Tanasa S, Shukla N, Cairo A, Ganji RS, Mikulkova P, Valuchova S, Raxwal VK, Capitao C, Schnittger A, Zdrahal Z et al (2023) A complex role of Arabidopsis CDKD;3 in meiotic progression and cytokinesis. Plant Direct 7:e477

Thazhath R, Liu C, Gaertig J (2002) Polyglycylation domain of beta-tubulin maintains axonemal architecture and affects cytokinesis in Tetrahymena. Nat Cell Biol 4:256–259

Torres MA, Nelson WJ (2000) Colocalization and redistribution of Dishevelled and actin during Wnt-induced mesenchymal morphogenesis. J Cell Biol 149:1433–1442

Tort O, Tanco S, Rocha C, Bieche I, Seixas C, Bosc C, Andrieux A, Moutin MJ, Aviles FX, Lorenzo J et al (2014) The cytosolic carboxypeptidases CCP2 and CCP3 catalyze posttranslational removal of acidic amino acids. Mol Biol Cell 25:3017–3027

Valenstein ML, Roll-Mecak A (2016) Graded control of microtubule severing by tubulin glutamylation. Cell 164:911–921

Vamadevan V, Chaudhary N, Maddika S (2022) Ubiquitin-assisted phase separation of Dishevelled-2 promotes Wnt signalling. J Cell Sci 135:jcs260284

van Dijk J, Miro J, Strub JM, Lacroix B, van Dorsselaer A, Edde B, Janke C (2008) Polyglutamylation is a post-translational modification with a broad range of substrates. J Biol Chem 283:3915–3922

van Dijk J, Rogowski K, Miro J, Lacroix B, Edde B, Janke C (2007) A targeted multienzyme mechanism for selective microtubule polyglutamylation. Mol Cell 26:437–448

Ververis A, Christodoulou A, Christoforou M, Kamilari C, Lederer CW, Santama N (2016) A novel family of katanin-like 2 protein isoforms (KATNAL2), interacting with nucleotide-binding proteins Nubp1 and Nubp2, are key regulators of different MT-based processes in mammalian cells. Cell Mol Life Sci 73:163–184

Wallingford JB, Habas R (2005) The developmental biology of Dishevelled: an enigmatic protein governing cell fate and cell polarity. Development 132:4421–4436

Wallingford JB, Rowning BA, Vogeli KM, Rothbacher U, Fraser SE, Harland RM (2000) Dishevelled controls cell polarity during Xenopus gastrulation. Nature 405:81–85

Wallkamm V, Rahm K, Schmoll J, Kaufmann LT, Brinkmann E, Schunk J, Kraft B, Wedlich D, Gradl D (2016) Regulation of distinct branches of the non-canonical Wnt-signaling network in Xenopus dorsal marginal zone explants. BMC Biol 14:55

Willsey HR, Walentek P, Exner CRT, Xu Y, Lane AB, Harland RM, Heald R, Santama N (2018) Katanin-like protein Katnal2 is required for ciliogenesis and brain development in Xenopus embryos. Dev Biol 442:276–287

Witte F, Bernatik O, Kirchner K, Masek J, Mahl A, Krejci P, Mundlos S, Schambony A, Bryja V, Stricker S (2010) Negative regulation of Wnt signaling mediated by CK1-phosphorylated Dishevelled via Ror2. FASEB J 24:2417–2426

Woodruff JB, Ferreira Gomes B, Widlund PO, Mahamid J, Honigmann A, Hyman AA (2017) The centrosome is a selective condensate that nucleates microtubules by concentrating tubulin. Cell 169:1066–1077.e1010

Wynshaw-Boris A (2012) Dishevelled: in vivo roles of a multifunctional gene family during development. Curr Top Dev Biol 101:213–235

Xie J, Han M, Zhang M, Deng H, Wu W (2018) PP5 (PPP5C) is a phosphatase of Dvl2. Sci Rep 8:2715

Yanagawa S, van Leeuwen F, Wodarz A, Klingensmith J, Nusse R (1995) The Dishevelled protein is modified by wingless signaling in Drosophila. Genes Dev 9:1087–1097

Ye B, Liu B, Hao L, Zhu X, Yang L, Wang S, Xia P, Du Y, Meng S, Huang G et al (2018) Klf4 glutamylation is required for cell reprogramming and early embryonic development in mice. Nat Commun 9:1261

Zhang K, Da Silva F, Seidl C, Wilsch-Brauninger M, Herbst J, Huttner WB, Niehrs C (2023) Primary cilia are WNT-transducing organelles whose biogenesis is controlled by a WNT-PP1 axis. Dev Cell 58:139–154.e138

Zwicker D, Decker M, Jaensch S, Hyman AA, Jülicher F (2014) Centrosomes are autocatalytic droplets of pericentriolar material organized by centrioles. Proc Natl Acad Sci USA 111:E2636–2645

Acknowledgements

Received funding from Czech Science Foundation grant GX19-28347X (VB); French National Research Agency grant ANR-10-IDEX-0001-02 (CJ); LabEx Cell'n'Scale grant ANR-11-LBX-0038 (CJ); Institut de convergence Q-life grant ANR-17-CONV-0005 (CJ); French National Research Agency grant ANR-12-BSV2-0007 (CJ); French National Research Agency grant ANR-17-CE13-0021 (CJ); Czech Academy of Sciences grant RVO: 86652036 (CB); Czech Science Foundation (Project GA22-25365S) (CB); Czech Science Foundation (Project GA23-07149S) (CB); Czech Science Foundation (Project GA22-06405S) (JH); and Charles University Grant Agency (project no. 1414120) (JN). The authors acknowledge the core facility CELLIM supported by the Czech-BioImaging large RI project (LM2018129 funded by MEYS CR) for their support with obtaining scientific data presented in this paper. CIISB research infrastructure project LM2018127 funded by MEYS CR and European Regional Development Fund-Project "UP CIISB" (No. CZ.02.1.01/0.0/0.0/18_046/0015974 are gratefully acknowledged for the financial support of the measurements at the Proteomics Core Facility (CEITEC) and at the Josef Dadok National NMR Centre. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures The project National Institute for Cancer Research (Programme EXCELES, ID Project No. LX22NPO5102) - Funded by the European Union – Next Generation EU. The authors would like to thank Bryjalab members and Mikołaj Nowak (Institut Curie) for technical support.

Author contributions

Marek Kravec: Conceptualization; Data curation; Formal analysis; Validation; Investigation; Visualization; Methodology; Writing—original draft;

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

 © The Author(s)

Writing—review and editing. **Ondrej Šedo**: Data curation; Software; Formal analysis; Investigation; Methodology; Writing—review and editing. **Jana Nedvědová**: Funding acquisition; Investigation; Writing—review and editing. **Miroslav Micka**: Software; Formal analysis; Investigation. **Marie Šulcová**: Formal analysis; Investigation; Visualization. **Nikodém Zezula**: Formal analysis; Investigation; Visualization. **Kristína Gömöryová**: Data curation; Software; Formal analysis; Visualization; Methodology; Writing—review and editing. **David Potěšil**: Formal analysis; Methodology. **Ranjani Sri Ganji**: Formal analysis. **Sara Bologna**: Resources. **Igor Červenka**: Conceptualization; Writing—review and editing. **Zbyněk Zdráhal**: Supervision; Writing—review and editing. **Jakub Harnoš**: Supervision; Methodology. **Konstantinos Tripsianes**: Supervision. **Carsten Janke**: Resources; Supervision; Funding acquisition; Writing—review and editing. **Cyril Bařinka**: Supervision; Funding acquisition; Writing—review and editing. **Vítězslav Bryja**: Conceptualization; Supervision; Funding acquisition; Methodology; Writing—original draft; Project administration; Writing—review and editing.

Source data underlying figure panels in this paper may have individual authorship assigned. Where available, figure panel/source data authorship is listed in the following database record: biostudies:S-SCDT-10_1038-S44318-024-00254-7.

Disclosure and competing interests statement

The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. Creative Commons Public Domain Dedication waiver http://creativecommons.org/publicdomain/zero/1.0/ applies to the data associated with this article, unless otherwise stated in a credit line to the data, but does not extend to the graphical or creative elements of illustrations, charts, or figures. This waiver removes legal barriers to the re-use and mining of research data. According to standard scholarly practice, it is recommended to provide appropriate citation and attribution whenever technically possible.

© The Author(s) 2024

Downloaded from https://www.embopress.org on October 1, 2024 from IP 147.251.22.14.

nature communications

Article

https://doi.org/10.1038/s41467-023-42503-z

Recognition and coacervation of G-quadruplexes by a multifunctional disordered region in RECQ4 helicase

Received: 9 February 2022

Accepted: 12 October 2023

Published online: 24 October 2023

Check for updates

Anna C. Papageorgiou[1,9], Michaela Pospisilova[2,3,9], Jakub Cibulka[3], Raghib Ashraf[2], Christopher A. Waudby[4,5], Pavel Kadeřávek[1], Volha Maroz[2,3], Karel Kubicek[1,2,6], Zbynek Prokop[7,8], Lumir Krejci[2,3,8] ✉ & Konstantinos Tripsianes[1] ✉

Biomolecular polyelectrolyte complexes can be formed between oppositely charged intrinsically disordered regions (IDRs) of proteins or between IDRs and nucleic acids. Highly charged IDRs are abundant in the nucleus, yet few have been functionally characterized. Here, we show that a positively charged IDR within the human ATP-dependent DNA helicase Q4 (RECQ4) forms coacervates with G-quadruplexes (G4s). We describe a three-step model of charge-driven coacervation by integrating equilibrium and kinetic binding data in a global numerical model. The oppositely charged IDR and G4 molecules form a complex in the solution that follows a rapid nucleation-growth mechanism leading to a dynamic equilibrium between dilute and condensed phases. We also discover a physical interaction with Replication Protein A (RPA) and demonstrate that the IDR can switch between the two extremes of the structural continuum of complexes. The structural, kinetic, and thermodynamic profile of its interactions revealed a dynamic disordered complex with nucleic acids and a static ordered complex with RPA protein. The two mutually exclusive binding modes suggest a regulatory role for the IDR in RECQ4 function by enabling molecular handoffs. Our study extends the functional repertoire of IDRs and demonstrates a role of polyelectrolyte complexes involved in G4 binding.

Guanine-rich nucleic acid sequences are prevalent in animal genomes and can spontaneously form G-quadruplexes (G4s) under physiological conditions[1]. These non-canonical secondary structures consist of repeating structural motifs called G-quartets that are stacked upon one another and are held together by an extensive network of hydrogen bonds linking four guanine bases around a cationic core[2–4]. G4s constitute a diverse family of highly stable DNA structures adopting parallel, hybrid or antiparallel topologies according to the pattern of the strand polarities and the orientation of the loops[2,5]. G4 sequences have been mapped within replication origins[6,7], mitochondrial DNA[8],

[1]CEITEC-Central European Institute of Technology, Masaryk University, Brno, Czech Republic. [2]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Brno, Czech Republic. [3]Department of Biology, Faculty of Medicine, Masaryk University, Brno, Czech Republic. [4]Institute of Structural and Molecular Biology, University College London, London WC1E 6BT, UK. [5]School of Pharmacy, University College London, London WC1N 1AX, UK. [6]Department of Condensed Matter Physics, Faculty of Science, Masaryk University, Brno, Czech Republic. [7]Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic. [8]International Clinical Research Center, St Anne's University Hospital, Brno, Czech Republic. [9]These authors contributed equally: Anna C. Papageorgiou, Michaela Pospisilova.
✉e-mail: lkrejci@chemi.muni.cz; kostas.tripsianes@ceitec.muni.cz

telomeric ends[9], gene promoter regions[10,11], and oncogenes[12], and therefore have been implicated in a range of biological processes. Although G4s act as functional regulatory elements in different cellular contexts, they often need to be unfolded to prevent impediment of various DNA metabolic processes[13] such as DNA replication[14–17], epigenetic control[18–20], telomere homeostasis[21], and transcription[22].

DNA helicases in the RECQ[23–26], FANCJ[27–29], DEAH/RHA[30,31], and PIF1[32,33] families prevent G4-induced genome instability by selectively recognizing and unwinding G4s[34]. Through scarce structural studies and FRET-based kinetics, G4 unwinding models have been proposed for a subset of these helicases[35–38]. Aside from DNA helicases, replication protein A (RPA), the most abundant ssDNA binding protein, has also been implicated in destabilizing G4 structures[39–41]. In particular, RPA has been shown to facilitate the role of several helicases by direct protein-protein interactions or trapping unfolded G4s[17,42,43].

Over the past years, protein phase separation has been reported in many biological processes, and nucleic acids are widely involved in regulating biological condensates[44]. For example, DEAD-box RNA helicases, play critical roles in many aspects of RNA metabolism by undergoing liquid-liquid phase separation (LLPS) with their RNA targets and other proteins[45]. However, LLPS studies on DNA helicases are still minimal[46], and it is not known if the formation of biomolecular condensates could be involved in G4 metabolism.

Among the RECQ DNA helicases, RECQ4 has a unique domain organization[47] (Fig. 1a). It lacks the RecQ C-terminal (RQC) domain required for G4 unwinding by the other family members. On the other hand, it features an N-terminal region that includes segments homologous to the Sld2 protein, an essential DNA replication factor in deeper-branching eukaryotes[48–51]. This Sld2-like region (amino acids 1–400) of RECQ4 is important for DNA replication[52] and is indispensable for viability in metazoans[53–55]. Biochemical analysis has revealed multiple DNA binding sites in the Sld2-like region, which can recognize various DNA structures with a preference for Holliday junctions[56] and G4s[57]. It also provides a platform for many interacting partners, suggesting that the Sld2-like region is integral to RECQ4 functions in DNA replication[58–62], DNA repair[61,63], and mitochondrial maintenance[64]. However, the Sld2-like region is largely unstructured[57,65] (Supplementary Fig. 1), raising the question of what mechanistic/functional aspects may be associated with the evolutionary conserved prevalence of intrinsic disorder[66–68].

Owing to their conformational flexibility and dynamics, intrinsically disordered regions (IDRs) expand the biomolecular functionality of ordered proteins and domains[69–72]. They act as hubs mediating signals upon interaction with multiple targets[73]; they adopt different structures upon binding to different partners[74,75]; they participate in ultra-high affinity, extreme disorder[76], multivalent[77–79], or polyelectrolyte[80] interactions with the ability to drive liquid-liquid phase separation[81–83] and the assembly of biomolecular condensates[84]. The diverse functions of IDRs are related to information encoded in their sequences[85]. For instance, IDRs of DNA binding proteins are rich in positively charged residues that engage in DNA contacts[86]. However, it is unknown if such disordered cationic segments can have alternative functions other than DNA binding and consequently may regulate biological processes.

Here, we identify a positively charged RECQ4-specific motif (RSM) that interacts with RPA by undergoing a disorder-to-helix transition in a static complex with RPA protein. The intrinsic disorder allows RSM to also engage in polyelectrolyte interactions with oppositely charged DNA molecules. Interestingly, binding to G4 structures is followed by associative phase separation. The comprehensive kinetic analysis revealed that the coacervation of G4 and RSM follows a rapid nucleation-growth mechanism resulting in a dynamic equilibrium between dilute and condensed phases. RSM thus can switch between phases, conformational states and binding partners, providing a mechanism allowing diverse RECQ4 functions in the cell.

Results

A RECQ4-specific motif mediates an interaction with RPA

Members of the RECQ family are known to interact with RPA by utilizing unstructured motifs in their sequences[87–91]. As it has been shown that RECQ4 depletion in *Xenopus* extracts suppresses RPA loading at the origin of replication[52], we tested for a physical interaction between the two proteins in human cells. Using an anti-GFP antibody, we carried out co-immunoprecipitation (co-IP) experiments of EGFP-tagged RECQ4 in U2OS cells. In contrast to the EGFP control, EGFP-RECQ4-WT was able to pull-down RPA, suggesting a possible direct interaction between the two proteins (Fig. 1b).

We then sought to validate the interaction between RECQ4 and RPA in vitro. RPA is a protein composed of three subunits (RPA70, RPA32, RPA14) and exhibits a modular architecture[92,93] that allows globular domains to associate in a trimeric core (70C, 32D, 14), bind ssDNA (70A, 70B, 70C, 32D), or participate in protein interactions (RPA70N, RPA32C) (Supplementary Fig. 2). Since RPA mainly associates with unstructured motifs present in BLM[88,90], WRN[87], and other binding partners[94–97], we reasoned that the largely disordered Sld2-like

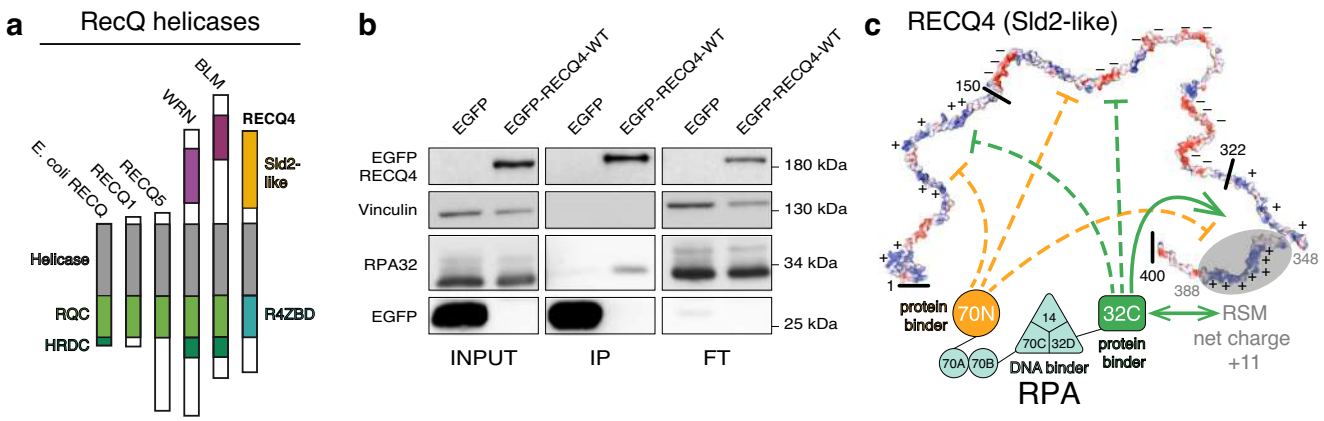


Fig. 1 | An interaction between RECQ4 and RPA. a Schematic representation of domain organization across the RecQ helicase family. **b** Immunoprecipitation of EGFP-RECQ4-WT or EGFP from whole cell lysates by GFP-trap beads. Samples were run on SDS-PAGE gel and immunoblotted on nitrocellulose membrane with indicated antibodies. The gel is a representative image of two independent experiments. **c** The topology of the RPA-RECQ4(Sld2-like) association was determined by NMR to detect physical interactions between segments. RPA modules are annotated as DNA or protein binders. For the RECQ4 fragment, RPA binding was further refined and located on RSM. Source data are provided as a Source Data file.

region of RECQ4 may contain the RPA-interacting site. Therefore, we performed a pull-down with the Sld2-like region of RECQ4 (aa 1–400) and RPA heterotrimer and confirmed the direct physical interaction (Supplementary Fig. 3a). The interaction was further characterized using 2D nuclear magnetic resonance (NMR) spectroscopy. The addition of unlabeled heterotrimeric RPA at a 2:1 molar ratio with ^{15}N-enriched RECQ4 (1–400) caused several peaks to broaden beyond detection (Supplementary Fig. 3b). Such line broadening of NMR signals results from a substantial increase in the overall molecular tumbling rate associated with the formation of a ~160 kDa RECQ4-RPA complex.

Next, we used an NMR-based approach to determine the minimal regions required for the interaction. The Sld2-like region of RECQ4 was divided into three polypeptides based on the charge distribution, while on the RPA side, the well-studied protein interaction modules RPA70N and RPA32C were selected[92,93] (Fig. 1c). 2D Heteronuclear single-quantum coherence (HSQC) spectra of ^{15}N enriched RECQ4 polypeptides were acquired in the absence and presence of a fourfold excess of each RPA domain. The fingerprint spectra indicate that the interaction occurs between the RECQ4 region spanning amino acids 322–400 and the RPA32C domain (Fig. 1c and Supplementary Fig. 3c, d). The specificity of the pairwise determination was corroborated by monitoring the chemical shift changes in RPA32C spectra induced by the addition of an excess of RECQ4$_{(322–400)}$ (Supplementary Fig. 4a).

To further delineate the region required for the interaction with RPA, we assigned the amino acid shifts of RECQ4$_{(322–400)}$ and classified them as perturbed or non-perturbed upon the addition of RPA32C (Supplementary Fig. 4a). This allowed us to map the RPA interaction within a RECQ4-specific motif (RSM; aa 348–388) that is conserved among various RECQ4 homologues (Supplementary Fig. 4b, c). RSM is intrinsically disordered and contains a basic patch with polyelectrolyte character composed of several positively charged residues (Figs. 1c and 2e). A complete set of NMR titration experiments for both RSM and RPA32C showed that the two proteins interact in a fast exchange regime in the NMR timescale (Fig. 2a, b). The larger chemical shift perturbations (CSPs) of RSM involved mainly the cationic residues of the positive patch, whereas those of RPA32C mapped to an acidic cleft on the surface of the domain (Fig. 2c, d). The NMR atomic-level view of the interaction is in excellent agreement with the common binding mode of RPA32C that utilizes the very same acidic cleft to associate with other partners, including UNG2, XPA, SMARCAL1, and TIPIN[94–96], all of them sharing positively charged residues in their interaction motif (Fig. 2d and Supplementary Fig. 5a). Unlike the other RPA partners, RECQ4 contains two well-conserved tryptophans in the binding motif. However, an RSM double tryptophan mutant (W379A/W383A) did not affect RPA32C binding in vitro as judged by the CSP binding profile (Supplementary Fig. 5b, c).

To ascertain the importance of the basic patch for the interaction with RPA, we reversed the charge in five amino acids (R375, K376, K380, K382, R384; 5E-mutant) based on the CSP quantification (Fig. 2e). As expected, the 5E-mutant abolished the binding to RPA32C in vitro (Supplementary Fig. 6). Next, we evaluated the interaction in a cell-based experiment by introducing doxycycline-inducible N-terminal EGFP-tagged RECQ4 mutant variant into U2OS cells (EGFP-RECQ4-5E) depleted for endogenous RECQ4 and performed co-immunoprecipitation experiments. We found that RPA co-precipitation is strongly attenuated with the RECQ4 5E-mutant

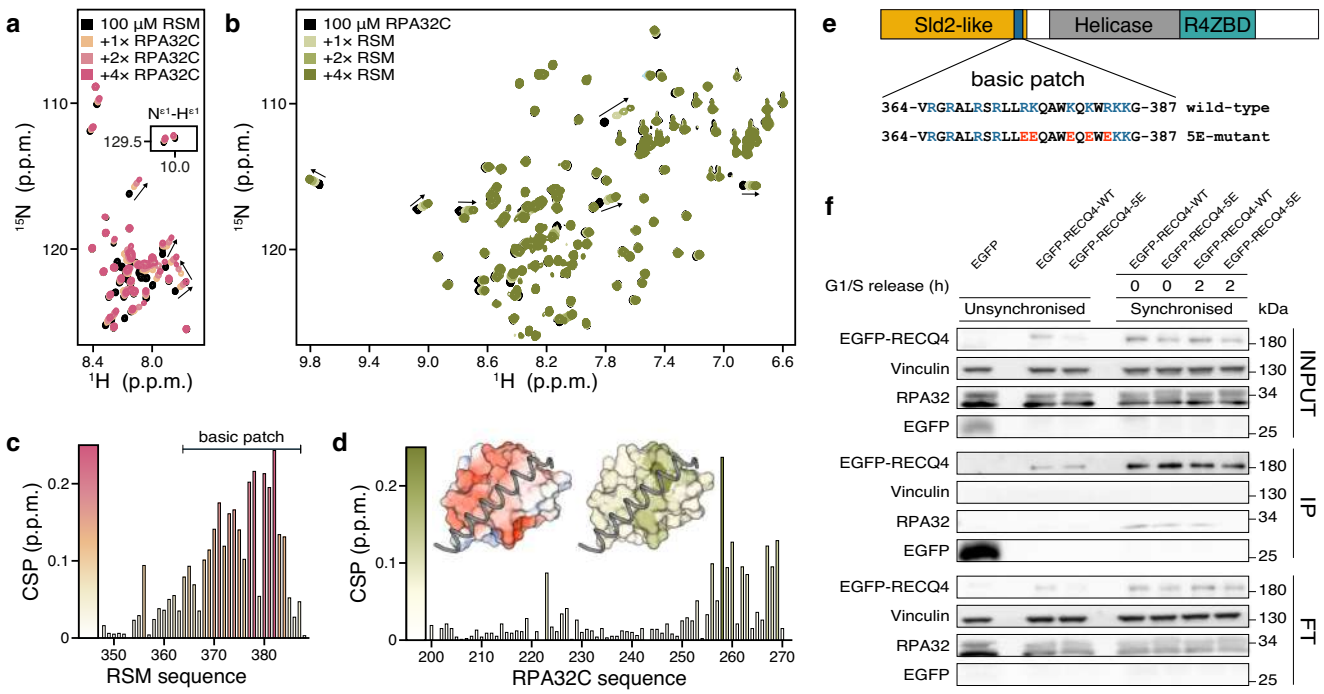


Fig. 2 | RSM binds the acidic cleft of RPA32C and is critical for RPA-binding in vivo. a, **b** Mapping the binary interaction between RSM and RPA32C by NMR titrations. ^{15}N labeled RSM titrated with zero to fourfold molar addition of RPA32C (**a**) and the reverse (**b**). Well-resolved chemical shift perturbations (CSP) are indicated with arrows. Inset in (**a**) shows the crosspeaks of tryptophan sidechains. **c**, **d** Per residue amide CSP of RSM (**c**) or RPA32C (**d**) induced by fourfold excess of the unlabeled partner. (**d**, inset) Crystal structure of RPA32C in complex with the SMARCAL1 peptide (PDB: 4mqv) showing (left) RPA electrostatics and (right) mapping of the RSM-induced CSPs. **e** Schematic depiction of RECQ4 protein showing the wild-type and 5E-mutant sequences of the basic patch. Positively charged residues are shown in blue, and charge reversal substitutions are in red. **f** Whole-cell lysates were immunoprecipitated from EGFP-RECQ4-WT, EGFP-RECQ4-5E, and EGFP unsynchronised or synchronised cells using GFP-trap beads. Bound proteins were separated on SDS-PAGE gel and immunoblotted with indicated antibodies on the nitrocellulose membrane. IP samples for RPA32 represent higher exposure of the membrane. The gel is a representative image of two independent experiments. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-023-42503-z

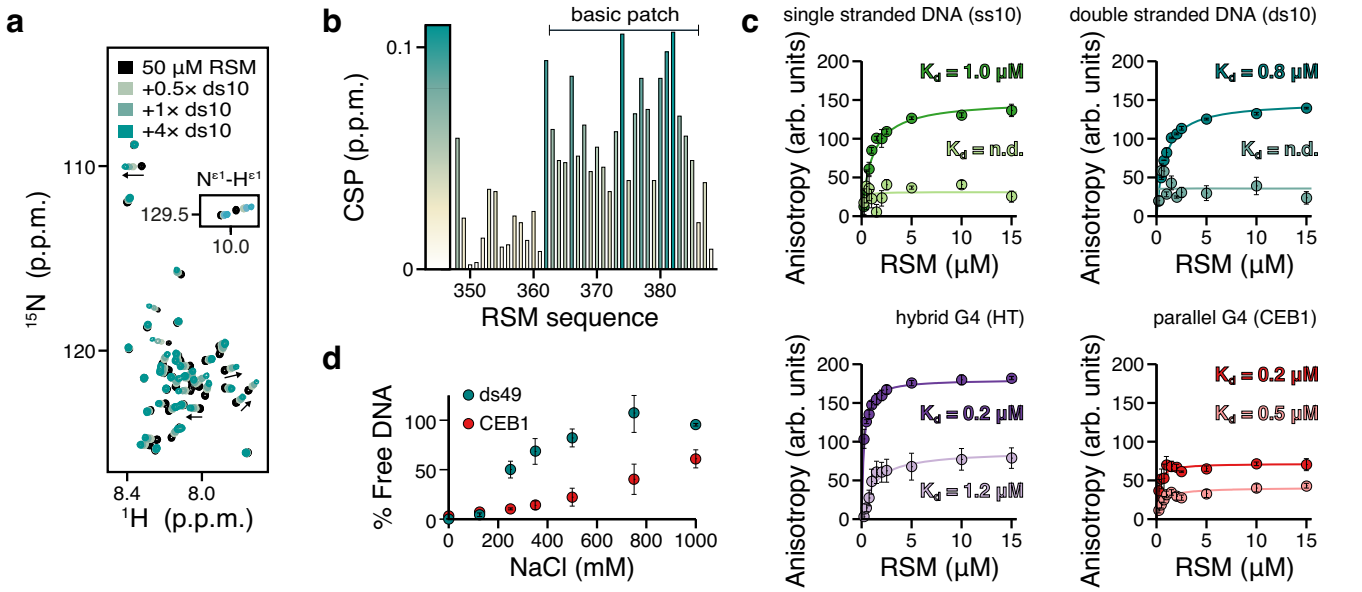


Fig. 3 | RSM forms electrostatically driven high-affinity complexes with DNAs. **a** ^{1}H-^{15}N HSQC spectra of 50 μM RSM titrated with double-stranded 10mer DNA (ds10). Some chemical shift perturbations (CSP) are indicated with arrows. Inset shows the crosspeaks of tryptophan sidechains. **b** CSPs of RSM residues induced by 4× molar addition of ds10. **c** Binding of RSM (dark color, 70 mM ionic strength; light color, 500 mM ionic strength) to ss10, ds10, HT, or CEB1 DNA monitored by FA measurements. $n = 3$ independent experiments; data are means ± s.d. **d** Effect of ionic strength on RSM binding to ds49 or CEB1 DNA quantified in EMSA experiments (see Supplementary Fig. 7f). $n = 3$ independent experiments; data are means ± s.d. Source data are provided as a Source Data file.

ex vivo, confirming that the basic patch of RSM is the principal determinant of the RPA-RECQ4 interaction (Fig. 2f).

RSM forms electrostatically driven high-affinity complexes with various DNA structures

The Sld2-like region of human RECQ4, particularly the $RECQ4_{(322–400)}$ fragment, has been previously shown to bind various DNA substrates with a strong preference for G4s[56,57,65,98]. To elucidate which part of $RECQ4_{(322–400)}$ possesses the DNA binding activity, we inspected the chemical shift changes of this fragment when bound to a 10-bp double-stranded DNA (ds10). The NMR analysis showed that only the residues within the RSM are affected by the presence of the DNA substrate (Supplementary Fig. 7a). To support this finding, we performed an Electrophoretic Mobility Shift Assay (EMSA) of various RECQ4 peptides within the 322–400 region with a 49-bp dsDNA (ds49) substrate. The complex binding profiles observed were indistinguishable, suggesting similar affinities for all tested peptides (Supplementary Fig. 7b, c). Our data thus demonstrate that the RSM itself or its shorter version obtained as a synthetic peptide (sRSM: 358–388) is sufficient for the DNA binding properties of the $RECQ4_{(322–400)}$ fragment[56].

Next, we monitored the interaction by NMR titrations to obtain atomic-level insight into the DNA binding properties. The RSM binds ds10 DNA in a fast exchange regime on the chemical shift timescale (Fig. 3a). CSP analysis showed that DNA binding is mediated by the positively charged residues of the basic patch (Fig. 3b), highlighting the dominant role of electrostatics in the interaction. We also compared the CSPs induced by other DNA substrates. The overall pattern was the same for a splayed-arm (Y14), single-stranded (ss10), or double-stranded DNA (ds10) with minor differences in the relative magnitude of CSPs (Supplementary Fig. 8). These data point strongly to polyelectrolyte interactions dominated by non-specific electrostatic contacts between the flexible cationic RSM and the negatively charged DNA molecules. Importantly, the binding of RSM to dsDNA did not affect the integrity of the DNA helix (Supplementary Fig. 7d). Interestingly, the CSP maps of the RSM bound to DNAs or RPA largely overlapped, suggesting that the RSM utilizes the same basic patch for different functions.

To examine the role of electrostatics in DNA binding, we performed Fluorescence Anisotropy (FA) experiments with ss10, ds10, and two G4 structures, one with hybrid (HT)[99] and one with parallel (CEB1)[100] topology (Supplementary Table 1). At low ionic strength, the binding affinities were in the submicromolar range for both G4s and around one micromolar for ss10 and ds10 DNAs (Fig. 3c). EMSA experiments further confirmed the RSM preference for G4 structures. In a mixed pool of equal amounts of CEB1 G4 and ss20 DNA, RSM bound the G4 structure with greater affinity than the ssDNA substrate (Supplementary Fig. 7e). At high ionic strength, binding assessed by FA was attenuated for the G4s and not reliably detected for both ss10 and ds10 DNAs (Fig. 3c). EMSA with increasing amounts of salt further supported the electrostatic nature of the DNA interactions. However, consistent with the FA results, RSM binding to CEB1 was more resistant to ionic strength than ds49 (Fig. 3d and Supplementary Fig. 7f). This data suggests that the versatile DNA binding properties of the RSM depend on ionic strength.

RSM conformational flexibility modulates distinct and competitive functions

Intrinsic disorder in proteins is often associated with promiscuous binding[101,102]. Indeed, our data demonstrate that RSM employs the physicochemical features of the basic patch to associate with RPA or various DNA substrates. To obtain detailed mechanistic information on RSM binding plasticity, we characterized the kinetic and thermodynamic parameters of the interactions along with structural and dynamic NMR descriptors.

A comparison of NMR titration measurements revealed that the relatively strong binding of RSM to dsDNA is associated with only small RSM chemical shift changes (CSPs not larger than 0.1 p.p.m., Fig. 3b). In contrast, the weaker binding of RSM to RPA32C is associated with much larger RSM chemical shift changes (CSPs as large as 0.25 p.p.m., Fig. 2c). These observations suggested that RSM complex formation with dsDNA or RPA32C may involve distinct binding mechanisms. To explore this, we compared the structural and dynamical traits of the RSM (Fig. 4a). Analysis of NMR secondary chemical shifts ($^{1}H^{\alpha}$, $^{1}H^{N}$, ^{15}N, $^{13}C'$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$) provides a sensitive indicator of dynamics and

https://doi.org/10.1038/s41467-023-42503-z

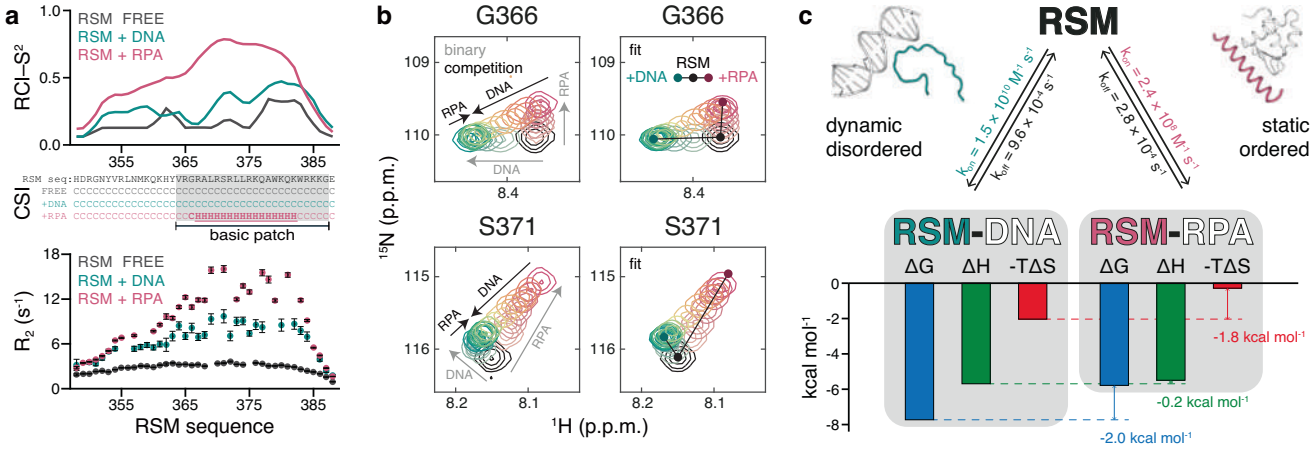


Fig. 4 | Distinct and competitive binding modes of RSM to DNA and RPA. **a** Random Coil Index order parameters (RCI-S^2) and secondary structure from Chemical Shift Index (CSI) (C: random coil, H: helix) calculated from secondary chemical shifts and ^{15}N R_2 relaxation rates measured for free RSM (black; 50 µM RSM) and RSM in complex with DNA (cyan; 4× molar excess dsDNA) or RPA (pink; 16× molar excess RPA32C). Relaxation experiments were performed once. Error bars represent the standard error of the fitted parameters. **b** Competitive binding of DNA (4× molar excess) and RPA (16× molar excess) to 50 µM RSM, observed by 2D NMR titrations and shown for two residues of RSM. Gray arrows indicate peak trajectories upon binding of each partner, and black arrows indicate peak trajectories upon competition with the other partner (left). Fitted spectra following 2D line shape analysis of the competition binding (right). **c** Summary of kinetic, thermodynamic, and structural parameters of RSM interactions. Source data are provided as a Source Data file.

structure[103–105]. The chemical shifts of free RSM indicated a highly disordered peptide with no intrinsic propensity for secondary structure formation (Fig. 4a and Supplementary Table 2). Similarly, the RSM chemical shifts in complex with dsDNA showed that it remains relatively flexible and unstructured (Fig. 4a and Supplementary Table 3). In contrast, the RSM chemical shifts in complex with RPA32C indicated increased order within the basic patch of RSM, associated with the formation of a stable α-helix (Fig. 4a and Supplementary Table 4). This is in line with studies of other peptides which undergo a disorder-to-helix transition upon RPA32C binding[94–96] (Fig. 2d and Supplementary Fig. 5b).

We also performed ^{15}N R_1, R_2, and {^1H}–^{15}N heteronuclear NOE (nuclear Overhauser effect) relaxation measurements to characterize the dynamic behavior of RSM (Fig. 4a and Supplementary Fig. 9a). The ^{15}N relaxation rates of free RSM are characteristic of a disordered peptide with no elements of increased rigidity. Binding to either dsDNA or RPA32C is mainly characterized by elevated R_2 values for the basic patch, which could be due to changes in rotational diffusion or due to slower processes like folding. Further analysis of the relaxation rates based on spectral density mapping[106,107] (Supplementary Fig. 9b, c) demonstrated internal rigidity for the basic patch in complex with dsDNA, which increased further when in complex with RPA32C. The relaxation data agree very well with the structural information encoded by the secondary chemical shifts and the entropy changes associated with binding (see below).

These NMR observations demonstrate two distinct modes of binding mediated by the RSM. RSM forms a dynamic complex with dsDNA, where internal motions are reduced, but the peptide remains flexible and disordered. In contrast, binding to RPA32C is accompanied by helix formation to yield a static ordered complex. However, both binding mechanisms rely on the positively charged RSM epitope, indicating that RPA32C and dsDNA interactions could be competitive and mutually exclusive.

To test this scenario and shed light on the multifunctional nature of the RSM, we performed 2D NMR competition measurements of ^{15}N-labeled RSM with dsDNA and RPA32C (Fig. 4b). First, RSM was titrated up to saturating quantities of dsDNA (4× excess) or RPA32C (16× excess). Then, each RSM complexed state was challenged with the competitor (Fig. 4b and Supplementary Fig. 10). In both competition experiments, RSM peaks lay along linear paths between the peak positions of the RPA32C- or dsDNA-bound states, confirming that RPA32C and dsDNA compete for RSM binding. Regardless of the order of the addition, the equilibrium established at the end was in favor of dsDNA binding due to the stronger affinity (Fig. 4b). To quantify these interactions, we performed a 2D lineshape analysis on the complete set of four titrations. This procedure simulates the complete set of 2D NMR experiments using a virtual spectrometer approach to determine the best-fitting spectral parameters (resonance frequencies and linewidths) and equilibrium and rate constants for a given binding model[108,109]. To determine dissociation constants and dissociation rates for dsDNA and RPA32C, RSM residues were fitted globally across 24 titration points to a competitive binding model (Fig. 4b, c). A good fit was obtained for all RSM residues, and the fitting quality did not improve by incorporating direct exchange between RPA32C- and dsDNA-bound states into the binding model. This analysis indicates that exchange between the two complexes occurs predominantly via dissociation and re-binding rather than direct displacement and agrees with the distinct structural states of the RSM when it associates with dsDNA or RPA32C.

The results of the 2D lineshape fitting above indicate that the affinity of RSM for dsDNA (K_d 6.3 ± 0.6 µM) is approximately 20-fold greater than for RPA32C (K_d 117 ± 5 µM). As dissociation rates are roughly comparable, the different affinities are driven primarily by the difference in the association rates, which we find to be $1.5 \pm 0.4 \times 10^{10}\,M^{-1}\,s^{-1}$ and $2.4 \pm 1.5 \times 10^{8}\,M^{-1}\,s^{-1}$ for dsDNA and RPA32C, respectively (Fig. 4c). Notably, the association rate for dsDNA exceeds even the simplest Smolochowski diffusion limit (ca. $10^9\,M^{-1}\,s^{-1}$, neglecting orientational constraints)[80,110], indicating that formation of the complex is steered by long-range electrostatic forces. However, given that the RPA32C binding cleft is also strongly negatively charged, the reduced association rate to RPA32C by two orders of magnitude is likely associated with the energetically costly formation of the observed α-helix. Therefore, we examined the interaction with dsDNA or RPA32C by Isothermal Titration Calorimetry (ITC) and analyzed the binding thermodynamics. The dissociation constants determined by ITC are in good agreement with the ones obtained from NMR competition experiments, confirming a 25-fold stronger binding affinity of RSM for dsDNA (Supplementary Fig. 11). The associated

https://doi.org/10.1038/s41467-023-42503-z

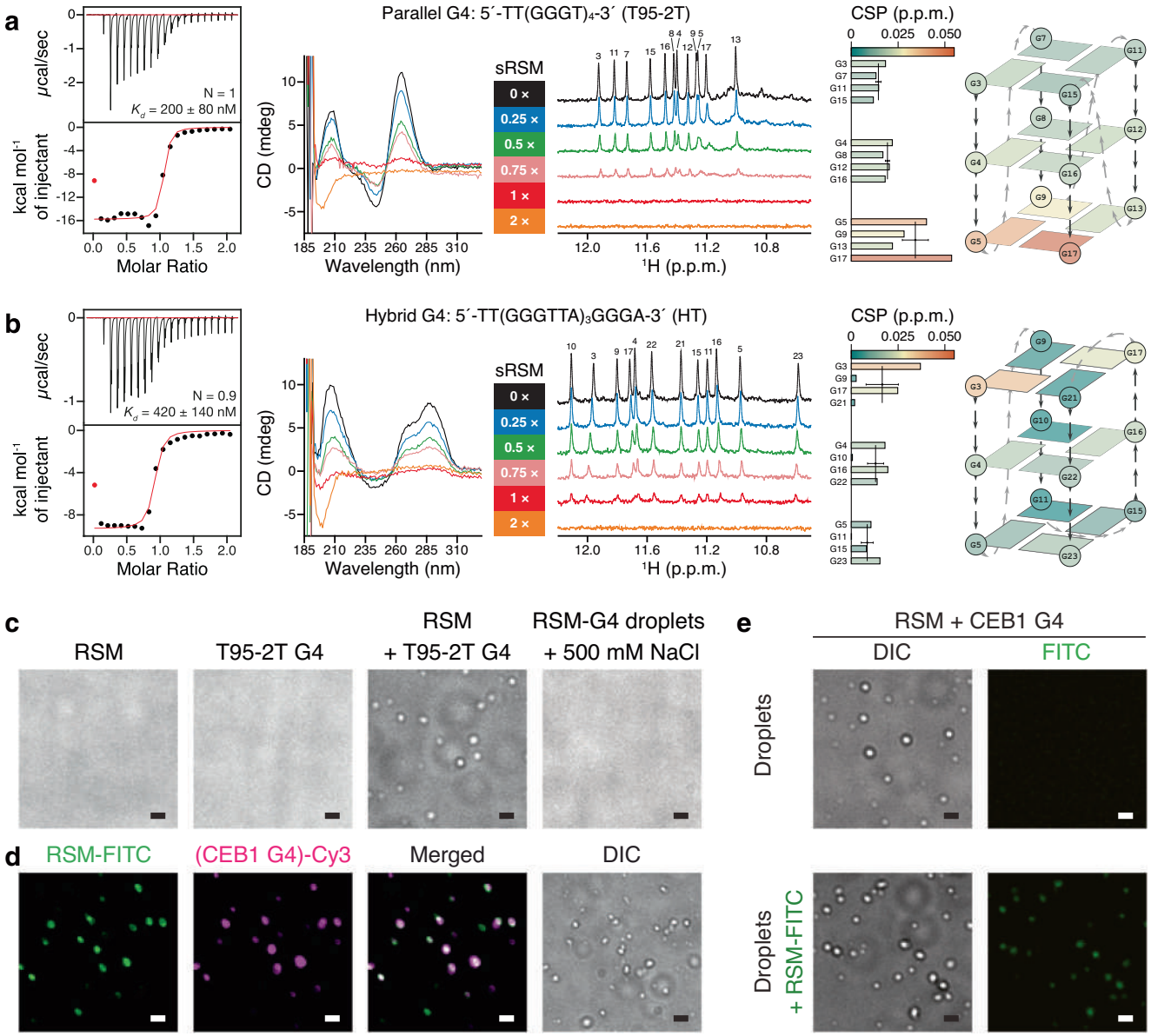


Fig. 5 | RSM-G4 recognition and coacervation. sRSM interaction with (**a**) parallel and (**b**) hybrid G4 was analyzed by ITC experiments (100 μM RSM), CD measurements (10 μM G4), or NMR titrations (50 μM G4). Imino chemical shift perturbations were quantified at sRSM:G4 ratio of 0.75:1 and color mapped in the schematic diagrams. G4 sequences are indicated. **c** RSM (10 μM), parallel T95-2T G4 (10 μM) or their mixture (10 μM each) were analyzed for droplet formation using DIC microscopy. The addition of 500 mM NaCl dissolves the droplets. Images are representative of three independent experiments. **d** Droplet formation and colocalization of 5 mol % FITC-labeled RSM (10 μM) and 5 mol % Cy3-labeled CEB1 G4 (10 μM). Images are representative of three independent experiments. **e** Droplet formation (upper panels) and colocalization (lower panels) of 5 mol% FITC-labeled RSM with preformed droplets. Images are representative of two independent experiments. In all images scale bar = 1 μm. Source data are provided as a Source Data file.

thermodynamic components (ΔG, ΔH, and ΔS) highlighted that the enthalpic contribution in both interactions is very similar, whereas substantial differences were observed in the entropic contribution (Fig. 4c). The large entropy difference cannot be attributed to hydrophobic interactions because the binding of both complexes involves the same positive patch of the RSM (Fig. 4b). Instead, the conformational properties of the RSM likely account for this phenomenon[111,112]. The dynamic nature of the RSM-dsDNA complex benefits from conformational flexibility, due to entropically favored interactions between any pairs of charged residues. In contrast, the formation of a static complex with RPA32C involves specific interactions and RSM folding into a well-defined helical configuration, which minimizes any entropic contribution. We conclude that the RSM conformational entropy modulates the kinetics and thermodynamics of binding to perform mutually exclusive functions.

Binding and coacervation between RSM and G-quadruplexes

Since RecQ family members (bacterial RECQ, WRN, BLM) are among the best-characterized helicases that unwind G4 motifs[113], we investigated the G4 binding properties of RSM in detail to understand RECQ4 roles in nucleic acid metabolism. For this, we used the T95-2T[114] and HT[99] G4s that fold into stable parallel and hybrid G4 topologies, respectively (Supplementary Fig. 12). The binding isotherms measured by ITC revealed that RSM binds to both quadruplexes with stoichiometry close to 1:1 and up to 8-fold higher affinity compared to dsDNA (dissociation constant of 200 ± 80 nM for the parallel and 420 ± 140 nM for the hybrid G4) (Fig. 5a, b and Supplementary Fig. 11). We also investigated RSM binding to G-quadruplexes by NMR. The RSM amide peaks in the HSQC spectra showed a dose-dependent reduction in their intensities upon the addition of G4s and completely disappeared at 1:1 stoichiometry (Supplementary Fig. 13a). Peak

intensities of positively charged residues were more reduced compared to the general trend (Supplementary Fig. 13b), pointing to the importance of the basic patch in G4 association. However, the fact that RSM peaks synchronously lose intensity (Supplementary Fig. 13a, b) is indicative of the possible formation of high-order complexes.

Indeed, the RECQ4 N-terminus is highly unstructured, with several regions predicted to undergo liquid-liquid phase separation (LLPS)[115]. Since the RSM shows the highest LLPS propensity (Supplementary Fig. 14a), we set out to analyze our samples for the formation of condensates by optical microscopy. While a solution of RSM or G4 alone was free of any visible particles, mixing RSM with G4 led to a rapid formation of droplets (Fig. 5c). Using fluorescently labeled components, we confirmed the presence of both RSM and G4 in the droplets (Fig. 5d). Formation of droplets was accompanied by sample turbidity that increased in a concentration-dependent manner (Supplementary Fig. 14b). We observed droplet formation with two types of parallel G4 (T95-2T and CEB1), hybrid G4 (HT), and ssDNA of varying length, with the notion that the shortest ssDNA showed very few droplets, but not with dsDNA or RPA32C (Supplementary Fig. 14d). This data indicate that RSM is prone to phase separation when mixed with oppositely charged polyelectrolytes, a process called complex coacervation that strongly depends on electrostatic interactions[116]. Indeed, RSM-DNA droplets are sensitive to salt concentration and dissolve above a critical salt concentration confirming the polyelectrolyte behavior of phase separation (Fig. 5c and Supplementary Fig. 14c). Likewise, the 5E-mutant that effectively neutralized the RSM charge did not form droplets when mixed with G4s. However, residual binding was still detectable by NMR between RSM-5E mutant and parallel G4 but not between RSM-5E mutant and hybrid G4 (Supplementary Fig. 14e, f). On the other hand, the RSM double tryptophan mutant was prone to forming droplets with both G4s, suggesting that the two tryptophan residues within the RSM are not critical for coacervation (Supplementary Fig. 14e, f).

To better understand the mechanism of RSM-G4 coacervation, we monitored the G4 proton resonances by NMR. The two selected G4 structures (T95-2T and HT) produced well-resolved imino spectra matching the assignments from the original reports and confirming G4 integrity and topology[99,114]. We performed NMR titrations by gradually adding RSM (aa 348–388) or sRSM (aa 358–388) to both G4s. The intensity of both imino and aromatic G4 peaks decreased proportionally with the dose of corresponding RECQ4 peptide to reach the noise level close to 1:1 stoichiometry (Fig. 5a, b and Supplementary Fig. 15). The progressive signal loss of G4 proton resonances reflects the droplet formation. Concurrently with the decrease of signal intensities, chemical shift perturbations were also observed, representing the first encounter between RSM and the G4 in the dilute solution phase on the path towards coacervation. Structural mapping of chemical shift perturbations in the parallel G4 suggests that the RSM engages first the 3′-end G-tetrad. In contrast, for the hybrid G4, the largest chemical shift perturbations are observed at the 5′-end G-tetrad, indicating that initial binding is sensitive to the G4 structure (Fig. 5a, b).

To further investigate the interaction between RSM and G4s, we performed circular dichroism (CD) measurements. Each G4 displays a unique CD spectral signature reflecting mainly G-quartet stacking, strand segment orientation, and loop arrangements[117]. Importantly, sRSM (aa 358–388) did not interfere with the G4 CD spectral features (Supplementary Fig. 16a). The G4-specific CD signal (at 265 nm for the parallel G4 and 286 nm for the hybrid G4) decreased in a dose-dependent fashion upon addition of synthetic RSM and disappeared at an equimolar ratio for both quadruplexes (Fig. 5a, b and Supplementary Fig. 16b). Since the CD observations correlated with the formation of droplets, we compared the CD spectra of samples in which the phase-separated sRSM-G4 were removed by centrifugation prior to measurement with those of samples containing the sRSM-G4 droplets. Interestingly, the effect on the magnitude of the CD spectrum was very similar regardless of the presence of droplets in the sample and proportional to the G4 amount remaining in the dilute phase collected after centrifugation (Supplementary Fig. 16c). It has been shown that the packing mode of dsDNA molecules in dispersion particles can change their optical properties[118]. Although it is unclear why RSM/sRSM-G4 droplets become optically inactive while other G4 coacervates do not[119], this unexpected finding presented the means to monitor the G4 equilibrium between dilute and condensed phases.

To gain a better understanding of the coacervation process between parallel G4 (T95-2T) and sRSM, a modern kinetic analysis was conducted, combining equilibrium titration data (Fig. 6a–c) with transient stopped-flow kinetics (Fig. 6d, e) in a global numerical model. Initially, a conventional analytical fitting was performed to demonstrate: (i) the presence of three distinct transitions between the dilute and condensed phases, (ii) a 1:1 stoichiometry, and (iii) a “cooperative” nature of sRSM-DNA assembly (Supplementary Fig. 17 and Supplementary Note). Subsequently, this information and the initial estimates of kinetic constants were used to develop a rigorous numerical model. The complex dataset, comprising two independent replicates for each experiment (Supplementary Fig. 18), was simultaneously fitted using numerical integration of the rate equations derived from the proposed model (Fig. 6f). The majority of the data were fitted in their raw form, except for the CD and NMR titrations which required additional processing to account for the concentration dependencies on the observed signals. Singular Value Decomposition (SVD) analysis of CD spectra concentration dependence demonstrated high similarity to the data acquired through a simple reading at two wavelengths (198 and 265 nm), indicating a straightforward relationship between the concentrations of sRSM and G4 and the resulting CD signal. Consequently, the specific-wavelength CD readings were utilized during the fit (Fig. 6a). On the contrary, the SVD analysis of NMR spectra provided valuable insights into the complex interplay between the reaction species and their influence on both the intensity and the chemical shift perturbation observed in the NMR titrations (Fig. 6b). Therefore, the complex NMR data were represented in the global fit through the SVD amplitude vectors.

Fitting the multiple data globally enabled the derivation of a unique set of rate constants for the individual steps of the proposed model for sRSM-G4 coacervation (Fig. 6f). The initial binding between sRSM and G4 occurred rapidly and reached equilibrium within 100 ms (Fig. 6d). The initial complex was relatively weak, with a dissociation constant ($K_{d,1}$) of $3.1 \pm 0.2\,\mu M$. The formation of the initial sRSM-G4 complex was slightly unfavorable, resulting in a positive free energy change ($\Delta G_{0,1}$) of $2.83\,kJ\,mol^{-1}$ (ΔG values are reported for a reference state of $1\,\mu M$). The second step of the coacervation process led to the formation of a stable nucleus ($n = 2$, the minimum value satisfying the model) with a dissociation constant ($K_{d,2}$) of 95 ± 7 nM. The free energy change ($\Delta G_{0,2}$) associated with nucleation is $-5.71\,kJ\,mol^{-1}$, strengthening the sRSM-G4 complex stability. The stable nucleus served as the foundation for subsequent growth ($n \geq 3$), which ultimately led to phase separation driven by a change in free energy ($\Delta G_{0,3}$) of $-3.88\,kJ\,mol^{-1}$ (Fig. 6g). The comprehensive kinetic analysis revealed that the coacervation of G4 and sRSM follows a rapid nucleation-growth mechanism. The simulation of the concentration of individual states (Fig. 6h) provided a visual representation of the dynamic equilibrium between the dilute and condensed phases under different conditions. The global numerical analysis provided a χ^2/DoF value of 1.12, indicating a good fit where an average global variance value is comparable to the average internal variance of the data. To further validate the accuracy of the model, we conducted spin-down assay experiments probing for the dilute species under varying concentrations. The experimental and simulated data (Fig. 6i and Supplementary Fig. 19) confirmed the consistency of the predictions and the reliability of the model in accurately capturing the dynamics of the system.

https://doi.org/10.1038/s41467-023-42503-z

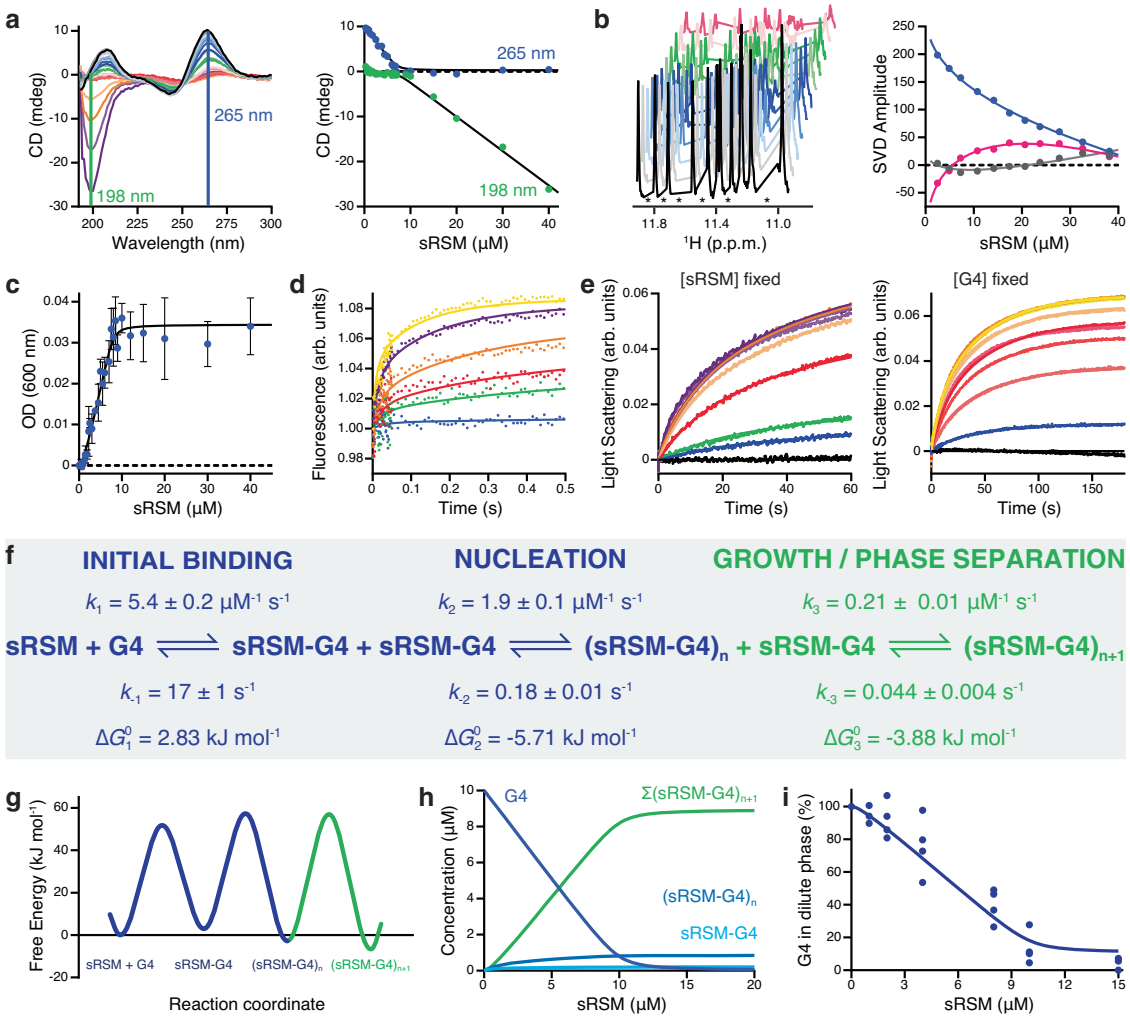


Fig. 6 | Kinetic model of sRSM-G4 coacervation. a CD spectra of 10 μM G4 titrated with 0–40 μM sRSM. The CD signal readings (right) at two specific wavelengths, 198 nm and 265 nm, were utilized in the global numerical analysis. **b** SVD amplitudes (right) calculated from the dependence of 50 μM G4 imino NMR spectra (left) on increasing sRSM concentration. Asterisks mark removed baseline signal regions to improve accuracy in SVD analysis. **c** The optical density at 600 nm monitored during the titration of 10 μM G4 with 0–40 μM sRSM. n = 3 independent experiments; data are means ± s.d. **d** The initial phase of the reaction examined by stopped-flow fluorescence upon mixing 2.5 μM G4 with 0–20 μM sRSM. **e** Kinetics of condensation process monitored by light scattering (at 295 nm) upon rapid mixing of 2.5 μM sRSM with 0–20 μM G4 (left) or 2.5 μM G4 with 0–20 μM sRSM (right). Each trace (in **d** and **e**) represents the average of 3 or 4 measurements. Solid lines in graphs from (**a**) to (**e**) represent the best global fit to the kinetic data. **f** The minimal kinetic model describing the sRSM-G4 coacervation includes three main steps: (i) initial binding, (ii) formation of stable nuclei and (iii) growth leading to phase separation. The rate constants for the individual steps were derived from global numerical fit. **g** Free energy profile of sRSM-G4 assembly and the reported free energy values in (**f**) were calculated using the Eyring equation (298 K; 1 μM RSM; 1 μM sRSM-G4). **h** The simulation represents the dynamic equilibrium of individual species for 10 μM G4 and sRSM concentration ranging from 0 to 20 μM. **i** The concentration of soluble species monitored using spin-down assay after incubation of 10 μM G4 with varying concentration of sRSM. Solid line represents the simulated data of G4 in dilute phase (sum of G4, sRSM-G4 and $(sRSM\text{-}G4)_2$ species), obtained from a global numerical model; data points from 4 experiments. Source data are provided as a Source Data file.

To demonstrate the dynamic nature of RSM-G4 coacervation, we performed further NMR experiments to perturb the final equilibrium. First, we introduced an excess of unlabeled sRSM to the stoichiometric mixture of ^{15}N RSM in complex with G4 that had already undergone phase-separation. This resulted in redistribution of the previously unobservable phase-separated ^{15}N RSM molecules between the dilute and condensed phases, resulting in the recovery of the NMR signal from the free ^{15}N RSM population in the dilute phase (Supplementary Fig. 13a). Consistently, FITC-labeled RSM was able to enter preformed RSM-G4 droplets (Fig. 5e). In a complementary experiment, we added an extra equivalent of G4 to a stoichiometric sRSM-G4 mixture that had formed droplets and showed no proton signals (Supplementary Fig. 13c). The resulting imino signals of excess G4 showed the sRSM induced perturbations, indicating a weighted average between the free and sRSM-bound populations in the dilute phase (Supplementary Fig. 13c and Fig. 5a). Additionally, the extra G4 resulted in the reappearance of sRSM methyl signals in the NMR spectrum, because of increased sRSM fraction (free or in complex with G4) in the dilute phase (Supplementary Fig. 13c). These data demonstrate further the exchange of sRSM and G4 between the solution and droplets.

To assess the G4 structural state across the phase equilibrium, we took advantage of Thioflavin T, a fluorescent G4-specific probe. Thioflavin T fluorescence is markedly increased upon binding to various G4s compared to ssDNA[120,121]. Indeed, we could observe a substantial increase in Thioflavin T fluorescence when incubated with T95-2T G4 but not with ssDNA (Fig. 7a, b). Intriguingly, a similar increase in fluorescence was observed upon induction of coacervation by adding RSM (Fig. 7a, b). These findings were further substantiated by visualizing Thioflavin T directly in the droplets by fluorescence microscopy (Fig. 7c, d). Although we can not exclude the possibility that ThT signal

https://doi.org/10.1038/s41467-023-42503-z

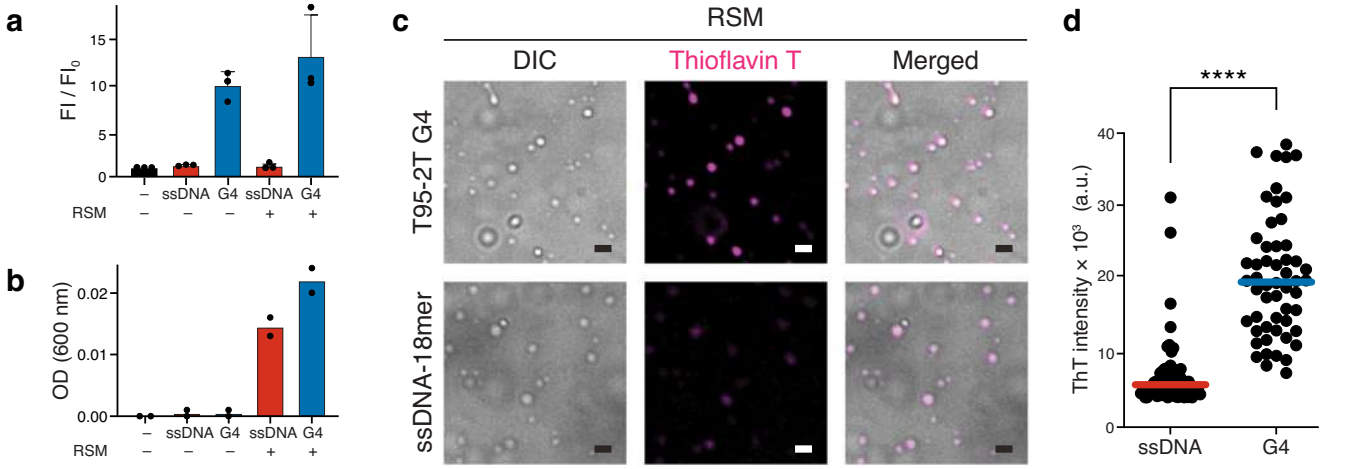


Fig. 7 | Analysis of G4 structure inside the droplets using Thioflavin T probe. **a** Thioflavin T (ThT, 500 nM) was mixed with T95-2T G4 (10 μM) or ssDNA−18mer (10 μM) in the absence or presence of RSM (10 μM) and the bulk fluorescence intensity was quantified. $n = 3$ independent measurements; data are means ± s.d. **b** Turbidity measurements of samples used for ThT fluorescence quantification in (**a**). Turbidity was measured as light absorbance at 600 nm. $n = 2$ independent measurements. **c** Phase separation microscopy of RSM (10 μM) mixed with T95-2T G4 or ssDNA−18mer (10 μM). ThT was added to a final concentration of 500 nM and the ThT fluorescence signal in the droplets was analyzed by fluorescent microscopy. In all images scale bar corresponds to 1 μm. **d** Manual quantification of ThT intensity in individual droplets from images in (**c**), $n = 54$ droplets for G4 and $n = 68$ droplets for ssDNA; $^{****}p < 0.0001$ (two-tailed Mann−Whitney test). Source data are provided as a Source Data file.

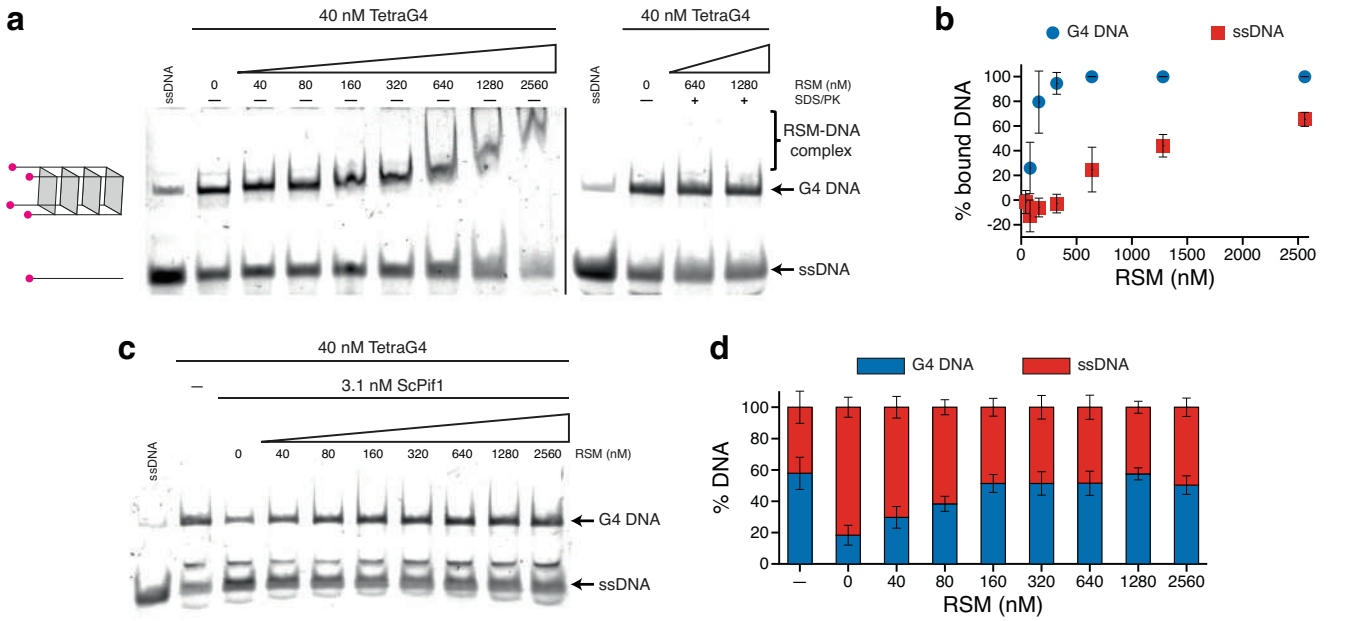


Fig. 8 | RSM binding to a tetramolecular G4 hinders G4 processing by helicases. **a** Binding of increasing RSM concentrations to 40 nM fluorescently labeled tetramolecular G4 (TetraG4) and ssDNA. Formation of RSM-G4 complex is reversed by addition of SDS/proteinase K. Schematic representation of ssDNA and folded G4 is shown on the left. **b** Quantification of gel image shown in (**a**); $n = 3$ independent experiments; data are means ± s.d. **c** ScPif1 helicase (3.1 nM) effectively unwinds tetramolecular G4 (40 nM), but its helicase activity is inhibited by pre-incubation of TetraG4 with RSM. **d** Quantification of gel image shown in (**c**), $n = 4$ independent experiments; data are means ± s.d. Source data are provided as a Source Data file.

may partially result from RSM amyloid structures formed in the droplets[122], our data suggest that G4 retains its molecular structure in the coacervate phase.

Searching for an RSM function related to G4 binding, we assembled a tetramolecular G4 structure previously used to monitor helicase unwinding activity[123]. CD measurements and binding assay with BG4 antibody verified tetramolecular G4 formation (Supplementary Fig. 20a, b). The addition of RSM did not result in G4 destabilization in contrast to unwinding by yeast Pif1 helicase (Supplementary Fig. 20c, d). Yet, EMSA verified the ability of RSM to bind tetramolecular G4 (Fig. 8a, b). We also tested the effect of RSM binding to G4 on Pif1 activity and observed robust inhibition of the helicase activity (Fig. 8c, d), indicating that the RSM-G4 complex represents a barrier for G4 unwinding. Similar inhibition of Pif1 was detected using $RECQ4_{(1-400)}$ (Supplementary Fig. 21a), confirming the RSM blocking activity in the context of the Sld2-like region. This inhibition was not limited to Pif1 helicase, as RSM also blocks G4 unwinding by another helicase, FANCJ (Supplementary Fig. 21b). Taken together, our data show that RSM exhibits high affinity for G4, which hinders G4 processing. However, the RSM-mediated inhibition of G4 unwinding is not solely a

consequence of associative phase separation, as we did not detect droplet formation by light microscopy under the conditions used in these experiments. In addition, a fusion of RECQ4$_{(322–400)}$ to APEX2 tag prevented visible phase separation in higher G4 concentrations (10 µM) but did not affect G4 binding and inhibition of G4 unwinding in lower G4 concentrations (40 nM) (Supplementary Fig. 22). These results highlight the specific interaction between RSM and G4, leading to the formation of a stable complex that interferes with the activity of helicases involved in G4 resolution.

Discussion

We have identified a positively charged region in human RECQ4 capable of forming polyelectrolyte complexes with oppositely charged nucleic acids. Interestingly, the DNA molecule profoundly affects the coacervate phase behavior of RSM-DNA complexes. G-quadruplexes exhibit a strong tendency to coacervate with RSM, whereas RSM-dsDNA complexes were not incorporated into droplets. RSM coacervation with ssDNA seems to depend on DNA length, possibly reflecting the total charge of the DNA molecule. However, the different macroscopic behaviors observed are not total-charge dependent because ds10, ss18, and G4 molecules (T95-2T and HT) have a similar net charge of around −20. Recent studies suggest that the complex coacervation of linear polypeptides is governed by charge patterning and/or charge density[124,125]. In the case of nucleic acids, molecular geometry affects the surface-charge patterning. For example, the disordered region of histone linker H1 promoted coacervation by increasing the charge density of the DNA via structuring[119]. These studies correlated rigidity and charge density of ssDNA, dsDNA, and G4 molecules having the same number of charges and ability to form droplets with H1[119]. The behavior of coacervates formed between a polylysine peptide and ssDNA or dsDNA demonstrated a similar effect depending on the stiffness and charge density of the DNA molecule[126]. Our observations on RSM-DNA coacervation differ markedly in that complexation with dsDNA did not result in droplet formation. This suggests that the electrostatic interactions that drive complex coacervation are not only influenced by the DNA charge distribution but also by the sequence (and charge) patterning of the disordered cationic peptide[85].

IDRs are characterized by high conformational entropy used to regulate their functions and interactions[127]. Our analysis demonstrates that RSM can switch between conformational states and perform diverse functions. On one hand, RSM remains highly flexible in polyelectrolyte interactions with nucleic acids. The disordered state of RSM in these complexes minimizes the entropic penalty of binding, resulting in relatively strong but non-specific interactions. On the other hand, RSM forms a stable helix when it interacts with RPA protein. The induced folding comes at an entropic cost, resulting in an interaction characterized by high specificity but low affinity. NMR competition experiments show that the two binding states are mutually exclusive. Although polyelectrolyte interactions with nucleic acids are favored over RPA32C binding in vitro, the spatiotemporal abundance of RPA molecules in the cell could trigger the transition between the disordered and ordered states of RSM. In other words, the conformational entropy of RSM can be modulated to rapidly regulate or coordinate sequential functions in cellular processes involving the RECQ4 helicase. However, we cannot exclude the possibility that the RSM-RPA32C interaction may be stabilized by other regions in the context of the RPA heterotrimer.

It is now recognized that protein IDRs are mechanistically involved in a wide range of cellular processes[128]. Our study extends this knowledge on the functions of IDRs, as it describes G4 binding within an IDR that hinders G4 processing by other helicases. RSM charge and inherent flexibility allow it to engage in strong electrostatic interactions with oppositely charged polyelectrolytes, such as G4 structures. Once in the polyelectrolyte complex, the G4 structure becomes inaccessible to processing by other helicases. Access to G4 is blocked due to RSM-G4 complex formation in solution and not because of subsequent coacervation observed in vitro. Based on our data, we speculate that the engagement of the disordered RSM with G4 represents a barrier for processing enzymes implying a possible signaling function of RECQ4 during temporal activation of replication origins[52,54,129]. Such a function may also provide mechanistic insights into the key role of RECQ4 in telomere[130,131] and mitochondrial[132,133] DNA maintenance or repair of DNA double-strand breaks[134–136] due to their direct G4 relevance[13,137].

Interestingly, these RSM properties seem to be evolutionary conserved since the *C. elegans* SLD2 protein, which shares the RSM physicochemical features, also forms droplets with G4 structures in vitro (Supplementary Fig. 23). However, a fusion of RECQ4$_{(322–400)}$ to APEX2, while not affecting G4 binding, eliminated charge-driven phase separation. This implies that the process may be regulated by structured domains within full-length RECQ4. With these considerations in mind, the great challenge ahead is to establish the localization of RECQ4-G4 complexes in phase-separated structures in the cell and the functional consequences[138].

In summary, our study uncovered an IDR in human RECQ4 with polyelectrolyte character. The physicochemical properties of the IDR account for the conformational plasticity in diverse functions and coacervation with oppositely charged polyelectrolytes such as G4 structures. Our findings offer interesting perspectives in understanding RECQ4 function in the cell linked to the partitioning of proteins into condensates.

Methods

Cell Culture

U2OS Flp-In T-Rex cell line (obtained from MRC PPU Reagents, University of Dundee in 2018) was maintained in DMEM media (LM-D1108/500; Biosera) with 10% tetracycline-free FBS (FB-1001T/500; Biosera) supplemented with Glutamine (XC-T1715/100; Biosera) and penicillin-streptomycin (XC-A412/100; Biosera) along with 100 µg/mL hygromycin (10687010; Invitrogen).

Stable cell line generation and cloning

To create EGFP-RECQ4-WT plasmid, RECQ4 DNA having EGFP and HA tag at N- and C-terminus, respectively, were cloned in the EcoRV sites of pAIO plasmid[139] (vector map is shown in the Source Data of the article) with a doxycycline-inducible system containing the modified promoter of medium strength to match the endogenous expression level of RECQ4. The plasmid also contains the shRNA target sequence (5′-TAGGAAGAGCCTCATCTAAG-3′) cloned between BglII and HindIII sites. Codon optimization was done against shRNA and siRNA in the RECQ4 CDS sequence. siRNA-RECQ4 S17991 (4392420, Thermo Fisher Scientific) was used to ensure maximum RECQ4 depletion with lipofectamine RNAiMAX (13778075, Invitrogen) for 48 h. To generate RECQ4 mutant (5E) or EGFP control cells, the corresponding mutations or stop codon respectively was introduced in the EGFP-RECQ4-WT plasmid by site-directed mutagenesis. Finally, stable cell lines were generated, and individual plasmids (500 ng) were transfected into U2OS Flp-In T-Rex cells (MRC PPU Reagents; University of Dundee) using Lipofectamine™ 3000 Transfection Reagent (L3000001; Invitrogen) along with Flp recombinase expression vector (1.5 µg) according to manufacturer's protocol and were selected with 100 µg/mL hygromycin. Expression of RECQ4 constructs was induced by 1 µg/mL Doxycycline (D9891; Sigma-Aldrich) for 48 h.

Co-immunoprecipitation

Immunoprecipitation was performed using GFP-Magnetic Agarose beads from a GFP-Trap Magnetic Agarose kit (gtmak-20; Chromotek) using 500 µg of whole cell lysate. Briefly, the cells were lysed in lysis buffer (provided by the IP kit), and the lysate was incubated with beads

for 1 h on a rotator at 4 °C, and 10% of lysate was kept as input. After incubation, the beads were separated on a magnetic separator, and 10% of the flow through was retained for analysis. The beads were washed three times with washing buffer, and the bound proteins were eluted by 2× SDS Laemmli buffer, heated at 100 °C for 5 min. All fractions were analyzed by SDS-PAGE, followed by western blotting with corresponding antibodies.

Synchronization of cells

Cells were synchronized by a double thymidine block. Cells were seeded, and thymidine (T1895; Sigma-Aldrich) was added to a final concentration of 2 mM for 16 h. After that, cells were released in fresh media without any thymidine for 6–8 h, followed by the new addition of thymidine for another 16 h and finally released in fresh media. Cells were then harvested according to the indicated time.

Western Blot

All samples were prepared in 2× SDS Laemmli buffer and separated on 10% SDS-PAGE at 100 V, followed by transfer of proteins to nitrocellulose membrane using the semi-dry Trans-blot turbo Transfer system (1704150; Biorad). After transfer, membranes were blocked in 5% milk/TBST for 1 h at room temperature and then incubated at 4 °C on a rocker overnight with the corresponding primary antibodies (Anti-RPA32, 1:700 dilution, A300-244A, Bethyl Laboratories; Anti-Vinculin, 1:5000 dilution, ab129002, Abcam; Anti-GFP, 1:2500 dilution, ab290, Abcam). The next day, the membranes were washed with TBST and incubated with the corresponding secondary antibodies (Anti-Rabbit IgG, 1:10,000 dilution, A6154, Sigma-Aldrich; Anti-Mouse IgG, 1:15,000, A0168, Sigma-Aldrich) for 1 h at room temperature. Finally, the blots were developed by the Immobilon Western Chemiluminescent HRP Substrate (WBKLS0500; MERCK Millipore), and images were acquired using the Luminescent Image Analyser (ImageQuant™ LAS 4000; Fujifilm).

DNA constructs and mutagenesis

Sequences of *H. sapiens* $RECQ4_{(1–150)}$, $RECQ4_{(150–315)}$, $RECQ4_{(322–400)}$, $RECQ4_{(348–388)}$ (RSM), and $RECQ4_{(1–400)}$ (UniProt code: O94761), *C. elegans* SLD2 protein (UniProt code: O44761), *H. sapiens* $RPA32_{(172–270)}$ (UniProt code: P15927), and *H. sapiens* $RPA70_{(1–120)}$ (UniProt code: P27694) were cloned in pET-M11 plasmid between NcoI and BamHI sites (vector map is shown in the Source Data of the article) to generate His_6 tag– TEV cleavage site fusions at the N-terminus of corresponding proteins.

DNA fragments were PCR amplified using Phusion high fidelity DNA polymerase master mix (M0530; NEB) and the primers listed in Supplementary Table 5, excised by restriction endonucleases (NcoI, R3193; BspHI, R0517; BamHI, R0136; NEB), and ligated using T4 DNA Ligase (M0202, NEB) to pET vector already treated with Antarctic Phosphatase (M0289; NEB). DNA sequences were confirmed by sequencing.

The RSM W379A/W383A and RSM 5E-mutants were generated by PCR-based site-directed mutagenesis (primers listed in Supplementary Table 5) using Pfu Turbo DNA polymerase (600250; Stratagene) and the $RECQ4_{(348–388)}$ expression vector as template. DNA sequences were confirmed by sequencing.

FLAG-APEX2-$RECQ4_{(322–400)}$ fragment was generated in an already existing plasmid pET-11d-FLAG-APEX2-RAD51. $RECQ4_{(322–400)}$, flanked by NcoI and SalI sites, was ligated into a vector digested by NcoI and XhoI, inserting a stop codon after RECQ4.

Protein preparation

Individual pET-M11 expression vectors (RECQ4, SLD2, and RPA) were transformed into *Escherichia coli* BL21(DE3) expression cells (C2527; NEB). Cells were grown at 37 °C in the presence of kanamycin. For ^{15}N- and ^{13}C-labeled samples, cells were grown in a minimal medium supplemented with $^{15}NH_4Cl$ (0.5 g/L) and $^{13}C_6$ glucose (2 g/L) as the sole nitrogen and carbon sources, respectively. Protein synthesis was induced by the addition of 0.5 mM of isopropyl-1-thio-d-galactopyranoside (A4773; Alchimica) at OD_{600} ~ 0.8, and cells were harvested after 5–6 h. Cells were resuspended in lysis buffer (25 mM Tris-HCl [pH 8], 400 mM NaCl, 20 mM imidazole, 2 mM beta-mercaptoethanol (BME), 0.1% (v/v) Triton X-100), lysed by sonication and centrifuged at 35,000 × *g*.

SLD2 and most RECQ4 proteins were purified in three steps. Initially, the lysate (denatured in 6 M urea) was applied to a HiTrap IMAC HP column charged with $CoCl_2$ (GE Healthcare) and eluted in buffer without urea supplemented with 500 mM imidazole. The protein-containing fractions were incubated at 4 °C with 0.5 mg TEV protease (produced in-house) per 1 L culture for 3 h at 25 °C to cleave the tag, followed by dialysis overnight at 4 °C against 25 mM Tris-HCl [pH 8], 100 mM NaCl, and 2 mM BME. The dialyzed protein sample was passed through a Mono S 4.6/100 PE column (GE HealthCare) and gradient eluted with 10 column volumes of buffer supplemented with 2 M NaCl. The protein-containing fractions were combined, concentrated and applied on HiLoad 10/300 Superdex Increase 75 GL column (GE HealthCare).

For the purification of $RECQ4_{(150–315)}$, RSM 5E-mutant, $RPA32_{(172–270)}$, and $RPA70_{(1–120)}$, the lysate (denatured for RECQ4 proteins and native for RPA proteins) was applied to a HiTrap IMAC HP column charged with $CoCl_2$ (GE Healthcare) and step eluted in buffer without urea supplemented with 500 mM imidazole. The protein-containing fractions were incubated at 4 °C with 0.5 mg of TEV protease (produced in-house) per 1 L culture for 3 h at 25 °C to cleave the tag and then dialyzed overnight at 4 °C against 25 mM Tris-HCl [pH 8], 400 mM NaCl, 10 mM imidazole, and 2 mM BME. The dialyzed protein samples were passed through the same HiTrap IMAC HP column. The non-bound protein was collected, concentrated and applied on HiLoad 10/300 Superdex Increase 75 GL column (GE HealthCare).

FLAG-APEX2-$RECQ4_{(322–400)}$ was expressed in in Rosetta 2 (DE3) pLysS (71403; Novagen) *E. coli* strain for 3.5 h at 37 °C. The cells were lysed by sonication in a lysis buffer containing 50 mM Tris [pH 7.5], 1 M KCl, 0.1 M sucrose, 1 mM EDTA, 0.01% NP40, 1 mM BME and protease inhibitors. Clarified lysate was loaded on a column with Affi-Gel Blue Gel matrix (Biorad) and eluted with a gradient of 0–2.5 M NaSCN. Fractions containing APEX2-$RECQ4_{(322–400)}$ were pooled and dialyzed against 20 mM K_2HPO_4 [pH 7.5], 10% glycerol, 0.5 mM EDTA, 200 mM KCl, 0.01% NP40 and 1 mM DTT, followed by another dialysis against 20 mM K_2HPO_4 pH 7.5, 10% glycerol, 0.5 mM EDTA, 100 mM KCl, 0.01% NP40 and 1 mM DTT. The dialyzed sample was loaded on a Mono S column (GE HealthCare), equilibrated with 20 mM K_2HPO_4 [pH 7.5], 10% glycerol, 0.5 mM EDTA, 100 mM KCl, 0.01% NP40 and 1 mM DTT and eluted with a gradient of 0.1–1 M KCl. Fractions containing pure protein were concentrated on the VivaSpin2 column with 30 MWCO.

The human RPA heterotrimer expression vector (RPA70-RPA32-RPA14, UniProt codes: P27694, P15927, P35244) was a kind gift from Marc S. Wold. The heterotrimeric RPA complex and MBP-$RECQ4_{(1–400)}$-His were expressed and purified as previously described[56,140].

Pull-down experiments

MBP-$RECQ4_{(1–400)}$-His and full-length RPA heterotrimer (8 µg of each) were incubated with Amylose Resin High Flow (E8022; NEB) for 30 min at 4 °C in 20 mM K_2HPO_4 [pH 7.5], 10% glycerol, 0.5 mM EDTA, 100 mM KCl, 0.01% NP40, and 2 mM BME. The unbound fraction was separated, the resin-bound fraction was washed, and fractions were analyzed by SDS-PAGE and Coomassie staining.

DNA substrates preparation

Used DNA substrates are listed in Supplementary Table 1. DNA oligos were purchased from Sigma or Generi Biotech. For NMR and CD

measurements, they were purified using a HiLoad 10/300 Superdex Increase 75 GL column (GE HealthCare) equilibrated with milliQ water. DNA-containing fractions were collected, lyophilized and resuspended in an annealing buffer containing 25 mM KPO4 [pH 6.5] and 70 mM KCl. For oligo annealing, initial heating at 95 °C for 10 min was followed by slow overnight cooling at room temperature. For phase separation experiments oligos were dissolved in 20 mM KPO4 [pH 7.0] and 70 mM KCl, heated at 100 °C for 5 min, and then rapidly cooled down by putting on ice. Tetramolecular quadruplex (TetraG4) was folded from the single DNA oligonucleotide in 20 mM Tris-HCl [pH 7.5], 1 mM EDTA, and 1 M NaCl as previously described[123]. Concentrations were determined by spectrophotometry using extinction coefficients provided by the manufacturer.

NMR spectroscopy

Experimental conditions (buffer composition and temperature) for NMR experiments shown in the manuscript are listed in Supplementary Table 6 according to the displayed item. For NMR and CD experiments shown in Fig. 5 that required high RSM stock concentration, a synthetic RSM peptide (sRSM: aa 358–388) was used (Peptide 2.0 Inc.). DNA binding and coacervation properties of RSM (aa 348–388) and sRSM (aa 358–388) did not differ (Supplementary Figs. 7b, c and 15a).

The $RECQ4_{(322–400)}$ sample used for chemical shift assignment contained ca. 0.7 mM, ^{13}C- and ^{15}N-labeled protein in 25 mM NaPO4 [pH 6.0], 100 mM NaCl, 2 mM BME, and 1 mM EDTA-d6 with 10% (v/v) ^2H_2O added for the lock. Spectra were recorded at 298 K using a 700 MHz Bruker Avance II spectrometer equipped with a TXO cryogenic probehead with *z* axis gradients. Protein backbone atom assignments were obtained using 3D CBCA(CO)NH, HNCACB, HNCA, HNCO, and HN(CA)CO experiments, and confirmed using a 4D HC(CC-TOCSY(CO))NH experiment[141,142]. The assignment completeness of $RECQ4_{(322–400)}$ exceeded 95% for all backbone chemical shifts (H^N, N, C^α, C^β, C′, H^α). The same set of 3D experiments was acquired along with a 3D HBHA(CBCACO)NH experiment to obtain the backbone RSM chemical shifts in complex with dsDNA or with RPA32C. NMR titrations of ^{15}N-labeled proteins (RECQ4 or RPA) were typically performed using ^1H-^{15}N HSQC experiments on samples containing 50 or 100 µM protein and step-wise addition of the binding partner. The assigned ^1H-^{15}N HSQC spectrum of RSM is shown in Supplementary Fig. 8. Normalized chemical shift perturbations (CSPs) were calculated by using the equation: $CSP = ([\delta_{HN}]^2 + [\delta_N/6]^2)^{0.5}$. RPA32C assignments were taken from BMRB entry 4460. Spectra were analyzed using SPARKY[143].

G4 samples (50 µM) were diluted in 25 mM KPO4 [pH 6.5] and 70 mM KCl in the absence or in the presence of increasing RSM molar equivalents. G4 and RSM were pre-incubated at 25 °C for 5 min before NMR acquisitions. NMR spectra were acquired at 278, 288, or 298 K on Bruker Avance spectrometers equipped with cryogenic probes operating in frequencies ranging from 600 to 950 MHz. 1D experiments were registered, processed, and analyzed using Topspin4.0 (Bruker Biospin).

Standard relaxation experiments were utilized for measurements of longitudinal and transverse relaxation rates[144]. The longitudinal relaxation rate experiments were measured with: 0.0111, 0.0555*, 0.1221, 0.1887, 0.2664, 0.3552, 0.4662*, 0.5883, 0.7548, 0.9657, 1.2876*, 1.665 s relaxation delays and the transverse relaxation rate experiments were measured with 0.0, 14.4*, 28.8, 43.2, 57.6, 72.0, 86.4*, 100.8, 115.2*, 144.0*, 172.8 ms relaxation delays (asterisks denote spectra measured twice in all cases). The longitudinal and transverse relaxation rates were obtained using a fit of the decays of signal intensities to a mono-exponential decay function $I_{i,n,k} = A_{i,n}\exp(-R_{i,n}\, t_{i,k})$ in program Octave 3.8.2[145] using the function *leasqr* from the package *optim*. The index i distinguishes variables related to the longitudinal ($i = 1$) and transverse relaxation ($i = 2$) rate measurement, index n stands for data of ^{15}N amide of an n-th residue, and index k is related to the measurement with the k-th relaxation delay. $I_{i,n,k}$ is the intensity of the signal read for a residue n in a spectrum acquired with the k-th relaxation delay of the length $t_{i,k}$. $A_{i,n}$ and $R_{i,n}$ are optimized parameters, where the former is a pre-exponential factor and $R_{i,n}$ is the relaxation rate for the n-th residue. 2000 Monte-Carlo simulations of the measured signal intensities were then generated, and the fit was repeated for each of them to estimate the error of the relaxation rates. The obtained relaxation rates were recalculated to spectral density values using reduced spectral density mapping protocol[146–148] with the assumption that the high-frequency values are similar and their difference can be neglected.

Steady state NOE experiments were measured with a 6 s saturation period composed of inversion pulses separated by 11.1 ms[149]. The saturation irradiation was replaced by a 20 s long interscan delay for the measurement of the reference spectra. NOE parameters were calculated for each amide peak as a ratio of signal intensities read in the spectrum with saturation (I_{sat}) and in the reference (I_{ref}) spectrum $NOE = I_{sat}/I_{ref}$. The errors of the signal intensities were estimated based on the noise in the spectra in the regions with no peaks. The errors of NOE were calculated using the error propagation law.

2D lineshape analysis of competitive binding

^1H-^{15}N HSQC titration spectra were fitted using TITAN v1.6[108]. In the latest release, a competitive binding model (now available, including source code) was created to describe the interaction between free RSM protein, P, and two competing ligands X and Y, representing RPA and DNA, to form the bound states B_X and B_Y:

$$P + X \underset{k^X_{off}}{\overset{k^X_d}{\rightleftharpoons}} B_X \tag{1}$$

$$P + Y \underset{k^Y_{off}}{\overset{k^Y_d}{\rightleftharpoons}} B_Y \tag{2}$$

These equilibria were solved numerically for given total protein and ligand concentrations and dissociation constants to determine the concentrations of individual species, from which the exchange superoperator, representing exchange between free and ligand-bound spin states, was formed:

$$\frac{d}{dt}\begin{pmatrix} P \\ B_X \\ B_Y \end{pmatrix} = \begin{pmatrix} -k^X_{on}[X] - k^Y_{on}[Y] & k^X_{off} & k^Y_{off} \\ k^X_{on}[X] & -k^X_{off} & 0 \\ k^Y_{on}[Y] & 0 & -k^Y_{off} \end{pmatrix} \cdot \begin{pmatrix} P \\ B_X \\ B_Y \end{pmatrix} \tag{3}$$

where $k^X_{on} = k^X_{off}/K^X_d$ (4) and $k^Y_{on} = k^Y_{off}/K^Y_d$ (5). This was subsequently incorporated into the TITAN simulation routine, allowing titration spectra to be fitted to determine the model parameters K^X_d, K^Y_d, k^X_{off} and k^Y_{off}. Four titration experiments were performed and combined to yield a total of 24 spectra that were then globally fitted, following standard protocols[109], using a selection of 16 representative residues. Reported uncertainties were determined using the jack-knife method.

Electrophoretic mobility shift assay

Cy3-labeled tetramolecular G4 (40 nM), Cy3-labeled CEB1 G4 (30 nM) and/or FITC-labeled ssDNA 20mer (30 nM) were incubated with increasing concentration of RSM in 40 mM Tris-HCl [pH 7.5], 25 mM KCl, 5 mM $MgCl_2$, 2 mM DTT, 2% glycerol, and 0.1 mg/mL BSA for 10 min at 37 °C. To assess the dependence of RSM binding on ionic strength, FITC-labeled CEB1 G4 (30 nM) or FITC-labeled ds49 (15 nM) was incubated with RSM (320 nM or 1280 nM) in the same buffer supplemented with the indicated concentration of NaCl for 10 min at 37 °C. To compare the binding of different RECQ4 fragments, we incubated FITC-labeled ds49 (15 nM) with increasing concentrations of

$RECQ4_{(322–400)}$, RSM or sRSM for 10 min at 37 °C. To verify the folding of G4, BG4 protein (produced in-house) was incubated with tetramolecular G4 (40 nM) in 50 mM Tris-HCl [pH 7.5], 50 mM KCl, and 1 mM $MgCl_2$ for 10 min at 37 °C followed by stabilization of BG4-G4 complex by 0.01% glutaraldehyde for another 5 min at 37 °C. After the addition of loading buffer (60% glycerol, 10 mM Tris-HCl [pH 7.5], and 60 mM EDTA), resulting complexes were analyzed either on 10% polyacrylamide gel in 1× TBE (90 mM Tris-HCl [pH 8], 9 mM boric acid, and 0.2 mM EDTA) for 1 h at 95 volts (for tetramolecular G4, ds49 and ionic strength) or on 0.8% agarose gel in 1× TBE supplemented with 10 mM KCl for 40 min at 70 V at 4 °C (for CEB1 G4 and ssDNA 20mer). Gels were scanned on a FLA-9000 scanner (Fujifilm) or Typhoon™ laser-scanner (Cytiva) and quantified with Multi Gauge version 3.2 (Fujifilm). For the quantification of ionic strength experiment, we included two additional controls of the DNA substrate alone in 750 mM and 1000 mM NaCl because the fluorescence intensity of the DNA substrate could be affected by increasing ionic strength. These controls were used to normalize the DNA fluorescence intensity of the corresponding NaCl concentration in the presence of RSM. DNA fluorescence intensity at the lower NaCl concentrations was normalized to the fluorescence intensity without any additional salt.

Fluorescence anisotropy measurements

Reactions containing 50 nM DNA substrate were incubated for 5–10 min at 25 °C with increasing concentrations of RSM and subsequently transferred to a 384-well microplate and read in a Tecan Microplate Reader Infinite F500 (Tecan group Ltd). Samples were excited with vertically polarized light at 490 nm, and both vertical and horizontal emissions were recorded at 525 nm. All measurements were conducted at 25 °C in 25 mM KPO4 [pH 6.5] and 70 mM KCl (low ionic strength) or 500 mM NaCl (high ionic strength). The data for each protein concentration was averaged over 10 min intervals to remove instrumental noise and processed by subtracting the anisotropy value obtained from the respective DNA substrate without the protein. Equilibrium dissociation constants (K_d) were calculated by fitting the data to the following equation:

$$FA = (([D] + [P] + K_d) - (([D] + [P] + K_d)^2 - (4\times[D]\times[P]))^{1/2})\times(A)/(2\times[D])$$

where [D] and [P] are concentrations of DNA and protein, respectively, and A is the maximum anisotropy value. Each data point is an average of three measurements.

Circular dichroism spectroscopy

All spectra were collected from 185 to 330 nm with a spectral bandwidth of 1 nm for each sample using a Chirascan™ CD Spectrometer at room temperature (~20–25 °C). An average of four scans was recorded. To examine changes in the G4 signal due to RSM binding, the G4 concentration was fixed at 10 µM, and sRSM (aa 358–388) was added to yield the mentioned molar ratios. The samples of parallel/hybrid G4 were prepared in 25 mM KPO4 [pH 6.5] and 70 mM KCl, and the spectra were measured after overnight incubation of samples at 4 °C (experiments shown in Fig. 5 and Supplementary Fig. 14) or 5 min of incubation at room temperature (experiments shown in Fig. 6 and Supplementary Fig. 22). To evaluate CD spectra, values of maxima (265 nm for parallel G4 and 198 nm for sRSM) were determined. Centrifuged samples were spun for 2 min at 13,400 rpm and the supernatant was collected for the measurements. The CD spectrum of tetramolecular G4 (8 µM) was recorded in 40 mM Tris-HCl [pH 7.5], 25 mM KCl, 5 mM $MgCl_2$, 2 mM DTT, 2% glycerol, and 0.1 mg/mL BSA. The CD spectrum of 10 µM single DNA oligonucleotide (non-folded control) used to assemble the tetramolecular G4 was recorded in water.

Isothermal titration calorimetry

Calorimetric titrations were carried out on an iTC200 microcalorimeter (MicroCal) at 30 °C. RSM and ligand (DNAs or RPA) samples were dissolved in 25 mM KPO4 [pH 6.5] and 70 mM KCl. The 300 µL sample cell was filled with a 0.1 mM solution of RSM, and the 40 µL injection syringe with 1 mM of the titrating ligand. Each titration consisted of a preliminary 0.2 µL injection followed by 19 subsequent 2 µL injections. The heat of the injections was corrected for the heat of dilution of every ligand into the buffer. Data fitting was done in Origin using the built-in one-site binding model.

Microscopy of RECQ4 droplets

RECQ4 fragments at 10 µM final concentration were mixed with an equimolar amount of indicated DNA substrate (or 126 ng/µL for plasmid DNA - pBlueScript) in 25 mM KPO4 [pH 6.5] and 70 mM KCl, and transferred to a chambered slide (Ibidi) for imaging. Images were acquired at room temperature using a DeltaVision Elite microscope (GE Healthcare) equipped with a 60× oil-immersion objective. Fluorescence images were deconvolved with SoftWoRx software (GE Healthcare) and all images were finally processed in ImageJ. The fluorescence intensity of Thioflavin T inside the droplets was quantified manually by creating a line profile in ImageJ through the centre of each droplet and recording the maximum intensity value.

Turbidity measurements

Increasing concentrations of RSM were mixed with 10 µM of the indicated DNA substrates in 25 mM KPO4 [pH 6.5] and 70 mM KCl, in a 96-well plate, and the optical density at 600 nm was measured on a Microplate Reader SpectraMax iD5 instrument immediately or after 5 min incubation at room temperature. For the ionic strength sensitivity experiment, increasing concentrations of NaCl (from 50 to 500 mM) were added to preformed droplets of sRSM-T95-2T and the optical density at 400 nm was measured.

Stopped-flow assay

Stopped-flow experiments were measured using an SFM-3000 stopped-flow machine (Bio-Logic) equipped with a MOS-500 monochromator spectrometer (Bio-Logic) and an additional Photomultiplier control unit (PMS-250) for light scattering analysis. The intrinsic fluorescence of tryptophan of sRSM was recorded with an excitation wavelength of 295 nm and 320 nm cut-off filter. Light scattering was measured as a signal reading from a photomultiplier placed at an angle of 90° to the light source (at 295 nm). All experiments were performed at 25 °C in 25 mM KPO4 [pH 6.5] and 70 mM KCl. In the experiments with a fixed concentration of G4, 2.5 µM of T95-2T G4 was mixed with various concentrations of sRSM, ranging from 0 to 20 µM. Conversely, in the experiments with a fixed concentration of sRSM, 2.5 µM of sRSM was systematically mixed with increasing concentrations of T95-2T G4, ranging from 0 to 20 µM. When sRSM was fixed, fluorescence intensity together with light scatter were collected for 60 s or 180 s in the first and second replicated measurement, respectively, according to the following protocol: (a) every 0.001 s from 0-1 s; (b) every 0.01 s from 1–10 s; (c) every 0.1 s from 10–180 s. When G4 was fixed, fluorescence intensity together with light scatter were collected for 60 s or 150 s in the first and second replicated measurement, respectively, according to the following protocol: (a) every 0.001 s from 0–0.3 s; (b) every 0.01 s from 0.3–10 s; (c) every 0.1 s from 10–180 s. The curves shown are averages of 3 to 4 technical replicates.

Spin-down assay

Indicated concentrations of sRSM and T95-2T G4 were mixed in 25 mM KPO4 [pH 6.5], 70 mM KCl and incubated at room temperature for 5 min. The samples were subsequently centrifuged for 3 min at 12,000 × *g* (room temperature), supernatant was collected and the pellet resuspended in an equal volume of buffer. Both fractions were deproteinized with 2.5 µL proteinase K (10 mg/ml) and 2.5 µL 1% SDS, run on an 0.8% agarose gel in 1× TBE supplemented with 10 mM KCl for 40 min at 70 volts at 4 °C and stained by SYBR® Gold Nucleic Acid Gel

Stain (S11494, Thermo Fisher Scientific, diluted 1: 10,000 in 1x TBE). The gels were scanned using Typhoon™ laser-scanner, quantified using Multi Gauge version 3.2 (Fujifilm).

Thioflavin T fluorescence quantification in solution

Thioflavin T (500 nM) was mixed with 10 µM of corresponding DNA substrates and where also indicated with 10 µM RSM in 25 mM KPO4 [pH 6.5] and 70 mM KCl. The fluorescence intensity at 490 nm (excitation wavelength: 445 nm) was measured by Microplate Reader SpectraMax iD5. The recorded values were normalized relative to the fluorescence of the Thioflavin T probe alone.

G4 unwinding assay

The complex of RSM, MBP-RECQ4$_{(1–400)}$-His or APEX2-RECQ4$_{(322–400)}$ with tetramolecular G4 was assembled by incubation at 37 °C for 10 min in 40 nM Tris-HCl [pH 7.5], 25 mM KCl, 5 mM $MgCl_2$, 2 mM DTT, 2% glycerol, and 0.1 mg/mL BSA supplemented with 1 mM ATP and ATP regeneration system (Creatine Phosphate-Creatine Kinase). The G4 unwinding was initiated by adding ScPif1 helicase (3.1 nM or 6.3 nM) or FANCJ helicase (2.5 nM). After incubation at 37 °C for 15 min, the reaction was stopped by adding a mixture of 1% SDS and 10 mg mL^{-1} proteinase K (A4392,0010, PanReac AppliChem) (1:1), resolved in 10% 1× TBE polyacrylamide gel at 95 volts for 1 h, and analyzed as above.

Kinetic data analysis and statistics

The kinetic data were fit globally with the KinTek Explorer program (KinTek, USA), a dynamic kinetic simulation program that allowed multiple data sets to be fit simultaneously to a single model. Data fitting used numerical integration of rate equations from an input model (Fig. 6f) searching a set of parameters using the Bulirsch–Stoer algorithm with an adaptive step size that produces a minimum χ^2 value calculated by using nonlinear regression based on the Levenberg-Marquardt method[150]. Residuals were normalized by sigma value for each data point.

The observable signal for circular dichroism data (CD) analyzed at 265 and 198 nm was defined as the sum of the contributions of G4 and RSM in soluble form (Equations 1 and 2), where a_{265} and a_{198} scales the signal to the concentration of G4 and sRSM-G4 at 265 and 198 nm, respectively and the factor b_{198} scales the signal to the concentration of sRSM and sRSM-G4 at 198 nm. The observable signal for optical density measurements (OD) was defined as an approximate contribution of condensation products (Eq. 3), where a_{OD} scales the signal to concentration and sensitivity of the measurement, the scaling factor b_{n+1} defines the relative change in optical density contribution of the products of continual growth (sRSM-G4)$_{n+1}$ ($n \geq 3$). Similarly, the observable signal for light scattering measurements (LS) was defined as an approximate contribution of condensation products ($n \geq 3$) (Eq. 4), where a_{LS} scales the concentration and b_{n+1} defines the relative change in optical density contribution of the consecutive condensation species (RSM-G4)$_{n+1}$ ($n \geq 2$). The observable fluorescence signal was defined as the sum of the contributions of each species to the total fluorescence (Eq. 5), where a_{FI} scales the signal to concentration and sensitivity of the measurement, factors b_{FI} and c_{FI} define the relative change in fluorescence in forming sRSM-G4 and (sRSM-G4)$_2$ complexes, respectively. Different scaling factors relating to the different sensitivity of analytical setups (factor a) were used for individual experiments, but the relative change in observable signal in forming individual complexes was constant for all datasets. The scaling factors were used as fitted parameters, and the obtained values are summarized in Supplementary Table 7.

$$Signal\left(CD_{265}\right) = a_{265}.\left(G4 + sRSM - G4\right) \tag{6}$$

$$Signal\left(CD_{196}\right) = a_{198}.\left(G4 + sRSM - G4\right) - b_{198}.\left(sRSM + sRSM - G4\right) \tag{7}$$

$$Signal\left(OD\right) = a_{OD}.\left(\left(sRSM - G4\right)_n + b_{n+1}.\left(sRSM - G4\right)_{n+1} + \ldots + b_i.\left(sRSM - G4\right)_i\right) \tag{8}$$

$$Signal\left(LS\right) = a_{LS}.\left(\left(sRSM - G4\right)_n + b_{n+1}.\left(sRSM - G4\right)_{n+1} + \ldots + b_i.\left(sRSM - G4\right)_i\right) \tag{9}$$

$$Signal\left(FI\right) = 1 + a_{FI}.sRSM - G4 + b_{FI}.\left(sRSM - G4\right)_2 \tag{10}$$

The concentration dependence of NMR spectra was resolved by singular value decomposition (SVD) analysis integrated within KinTek Explorer. Extracted SVD amplitude vectors were then fitted to a proposed kinetic model by nonlinear regression analysis based on the numerical integration of rate equations. Alternating between the fitting of SVD amplitude vectors and the direct fitting of spectra helped us find the best global fit to the data. To account for slight variations in the data, sRSM or G4 concentrations were allowed to vary within an interval of ±10% to make the best fits possible. The standard error (s.e.) was calculated from the covariance matrix during nonlinear regression. The standard error estimates in fitted parameters were propagated to yield error estimates in calculated values, the equilibrium dissociation constant (K_d).

The free energy of binding was calculated for a reference state rather than the 1 M standard state (Eq. 11), where the product $K.[L] = [L]/K_d$ is a dimensionless number that defines the thermodynamic driving force for binding at the physiological concentration of ligand concentration[151]. The free energy profile was calculated using the Eyring equation (Eq. 12) at reference temperature 298 K, and reference concentration of 1 µM sRSM and 1 µM sRSM-G4.

$$\Delta G_0 = -R.T.\ln(K.[L]) \tag{11}$$

$$\Delta G^{\ddagger} = -R.T.\ln(k.h/k_B.T) \tag{12}$$

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All the PDB codes cited in this paper (4MQV, 4GOP, 5N85, 2LK7, 2GKU) are available in the protein data bank web server. Chemical shift assignments of RPA32C were obtained from BMRB code 4460. The chemical shift assignments of RECQ4$_{(322–400)}$ can be accessed using the following accession code: BMRB 51487. All data generated in this study are available within the Article and Supplementary Information. Source data are provided with this paper.

References

1. Bochman, M. L., Paeschke, K. & Zakian, V. A. DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.* **13**, 770–780 (2012).
2. Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K. & Neidle, S. Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.* **34**, 5402–5415 (2006).
3. Williamson, J. R., Raghuraman, M. K. & Cech, T. R. Monovalent cation-induced structure of telomeric DNA: the G-quartet model. *Cell* **59**, 871–880 (1989).

4. Sfen, D. & Gilbert, W. Formation of parallel four-stranded complexes by guanine-rich motif for meiosis. *Nature* **334**, 364–366 (1998).
5. Spiegel, J., Adhikari, S. & Balasubramanian, S. The structure and function of DNA G-quadruplexes. *Trends Chem*. **2**, 123–136 (2020).
6. Prorok, P. et al. Involvement of G-quadruplex regions in mammalian replication origin activity. *Nat. Commun*. **10**, 3274 (2019).
7. Valton, A. L. et al. G4 motifs affect origin positioning and efficiency in two vertebrate replicators. *EMBO J*. **33**, 732–746 (2014).
8. Huang, W. C. et al. Direct evidence of mitochondrial G-quadruplex DNA by using fluorescent anti-cancer agents. *Nucleic Acids Res*. **43**, 10102–10113 (2015).
9. Koirala, D. et al. Intramolecular folding in three tandem guanine repeats of human telomeric DNA. *Chem. Commun*. **48**, 2006–2008, (2012).
10. Siddiqui-Jain, A., Grand, C. L., Bearss, D. J. & Hurley, L. H. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl Acad. Sci. USA* **99**, 11593–11598 (2002).
11. Balasubramanian, S., Hurley, L. H. & Neidle, S. Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nat. Rev. Drug Discov*. **10**, 261–275 (2011).
12. Yuan, L. et al. Existence of G-quadruplex structures in promoter region of oncogenes confirmed by G-quadruplex DNA cross-linking strategy. *Sci. Rep*. **3**, 1811 (2013).
13. Varshney, D., Spiegel, J., Zyner, K., Tannahill, D. & Balasubramanian, S. The regulation and functions of DNA and RNA G-quadruplexes. *Nat. Rev. Mol. Cell Biol*. **21**, 459–474 (2020).
14. Besnard, E. et al. Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nat. Struct. Mol. Biol*. **19**, 837–844 (2012).
15. Lerner, L. K. & Sale, J. E. Replication of G Quadruplex DNA. *Genes* **10**, 95–95 (2019).
16. Maffia, A., Ranise, C. & Sabbioneda, S. From R-loops to G-quadruplexes: emerging new threats for the replication fork. *Int. J. Mol. Sci*. **21**, 1506 (2020).
17. Lee, W. T. C. et al. Single-molecule imaging reveals replication fork coupled formation of G-quadruplex structures hinders local replication stress signaling. *Nat. Commun*. **12**, 2525 (2021).
18. Sarkies, P., Reams, C., Simpson, L. J. & Sale, J. E. Epigenetic instability due to defective replication of structured DNA. *Mol. Cell* **40**, 703–713 (2010).
19. Mao, S. Q. et al. DNA G-quadruplex structures mold the DNA methylome. *Nat. Struct. Mol. Biol*. **25**, 951–957 (2018).
20. Reina, C. & Cavalieri, V. Epigenetic modulation of chromatin states and gene expression by G-quadruplex structures. *Int. J. Mol. Sci*. **21**, 1–22 (2020).
21. Paeschke, K., Simonsson, T., Postberg, J., Rhodes, D. & Lipps, H. J. Telomere end-binding proteins control the formation of G-quadruplex DNA structures in vivo. *Nat. Struct. Mol. Biol*. **12**, 847–854 (2005).
22. Hansel-Hertsch, R. et al. G-quadruplex structures mark human regulatory chromatin. *Nat. Genet*. **48**, 1267–1272 (2016).
23. Sun, H., Karow, J. K., Hickson, I. D. & Maizels, N. The Bloom's syndrome helicase unwinds G4 DNA. *J. Biol. Chem*. **273**, 27587–27592 (1998).
24. Sun, H., Bennett, R. J. & Maizels, N. The Saccharomyces cerevisiae Sgs1 helicase efficiently unwinds G-G paired DNAs. *Nucleic Acids Res*. **27**, 1978–1984 (1999).
25. Fry, M. & Loeb, L. A. Human werner syndrome DNA helicase unwinds tetrahelical structures of the fragile X syndrome repeat sequence d(CGG)n. *J. Biol. Chem*. **274**, 12797–12802 (1999).
26. Wu, X. & Maizels, N. Substrate-specific inhibition of RecQ helicase. *Nucleic Acids Res*. **29**, 1765–1771 (2001).
27. Wu, Y., Shin-ya, K. & Brosh, R. M. Jr FANCJ helicase defective in Fanconia anemia and breast cancer unwinds G-quadruplex DNA to defend genomic stability. *Mol. Cell Biol*. **28**, 4116–4128 (2008).
28. Bharti, S. K. et al. Specialization among iron-sulfur cluster helicases to resolve G-quadruplex DNA structures that threaten genomic stability. *J. Biol. Chem*. **288**, 28217–28229 (2013).
29. Wu, C. G. & Spies, M. G-quadruplex recognition and remodeling by the FANCJ helicase. *Nucleic Acids Res*. **44**, 8742–8753 (2016).
30. Vaughn, J. P. et al. The DEXH protein product of the DHX36 gene is the major source of tetramolecular quadruplex G4-DNA resolving activity in HeLa cell lysates. *J. Biol. Chem*. **280**, 38117–38120 (2005).
31. Sauer, M. et al. DHX36 prevents the accumulation of translationally inactive mRNAs with G4-structures in untranslated regions. *Nat. Commun*. **10**, 2421 (2019).
32. Paeschke, K., Capra, J. A. & Zakian, V. A. DNA replication through G-quadruplex motifs is promoted by the Saccharomyces cerevisiae Pif1 DNA helicase. *Cell* **145**, 678–691 (2011).
33. Paeschke, K. et al. Pif1 family helicases suppress genome instability at G-quadruplex motifs. *Nature* **497**, 458–462 (2013).
34. Wu, L. & Hickson, I. D. DNA helicases required for homologous recombination and repair of damaged replication forks. *Annu. Rev. Genet*. **40**, 279–306 (2006).
35. Tippana, R., Hwang, H., Opresko, P. L., Bohr, V. A. & Myong, S. Single-molecule imaging reveals a common mechanism shared by G-quadruplex-resolving helicases. *Proc. Natl Acad. Sci. USA* **113**, 8448–8453 (2016).
36. Chen, W. F. et al. Molecular mechanistic insights into Drosophila DHX36-mediated G-quadruplex unfolding: a structure-based model. *Structure* **26**, 403–415.e404 (2018).
37. Chen, M. C. et al. Structural basis of G-quadruplex unfolding by the DEAH/RHA helicase DHX36. *Nature* **558**, 465–469 (2018).
38. Voter, A. F., Qiu, Y., Tippana, R., Myong, S. & Keck, J. L. A guanine-flipping and sequestration mechanism for G-quadruplex unwinding by RecQ helicases. *Nat. Commun*. **9**, 4201 (2018).
39. Qureshi, M. H., Ray, S., Sewell, A. L., Basu, S. & Balci, H. Replication protein A unfolds G-quadruplex structures with varying degrees of efficiency. *J. Phys. Chem. B* **116**, 5588–5594 (2012).
40. Safa, L. et al. Binding polarity of RPA to telomeric sequences and influence of G-quadruplex stability. *Biochimie* **103**, 80–88 (2014).
41. Safa, L. et al. 5′ to 3′ Unfolding directionality of DNA secondary structures by replication protein A: G-quadruplexes and duplexes. *J. Biol. Chem*. **291**, 21246–21256 (2016).
42. Wu, W. et al. HERC2 facilitates BLM and WRN helicase complex interaction with RPA to suppress G-quadruplex DNA. *Cancer Res*. **78**, 6371–6385 (2018).
43. Maestroni, L. et al. RPA and Pif1 cooperate to remove G-rich structures at both leading and lagging strand. *Cell Stress* **4**, 48–63 (2020).
44. Wang, B. et al. Liquid-liquid phase separation in human health and diseases. *Signal Transduct. Target Ther*. **6**, 290 (2021).
45. Weis, K. & Hondele, M. The role of DEAD-box ATPases in gene expression and the regulation of RNA-protein condensates. *Annu. Rev. Biochem*. **91**, 197–219 (2022).
46. Wang, T. et al. Bloom syndrome helicase compresses single-stranded DNA into phase-separated condensates. *Angew. Chem. Int. Ed. Engl*. **61**, e202209463 (2022).
47. Kaiser, S., Sauer, F. & Kisker, C. The structural and functional characterization of human RecQ4 reveals insights into its helicase mechanism. *Nat. Commun*. **8**, 15907 (2017).
48. Masumoto, H., Muramatsu, S., Kamimura, Y. & Araki, H. S-Cdk-dependent phosphorylation of Sld2 essential for chromosomal

DNA replication in budding yeast. *Nature* **415**, 651–655 (2002).
49. Tanaka, S. et al. CDK-dependent phosphorylation of Sld2 and Sld3 initiates DNA replication in budding yeast. *Nature* **445**, 328–332 (2007).
50. Zegerman, P. & Diffley, J. F. Phosphorylation of Sld2 and Sld3 by cyclin-dependent kinases promotes DNA replication in budding yeast. *Nature* **445**, 281–285 (2007).
51. Gaggioli, V. et al. CDK phosphorylation of SLD-2 is required for replication initiation and germline development in C. elegans. *J. Cell. Biol.* **204**, 507–522 (2014).
52. Sangrithi, M. N. et al. Initiation of DNA replication requires the RECQL4 protein mutated in Rothmund-Thomson syndrome. *Cell* **121**, 887–898 (2005).
53. Matsuno, K., Kumano, M., Kubota, Y., Hashimoto, Y. & Takisawa, H. The N-terminal noncatalytic region of Xenopus RecQ4 is required for chromatin binding of DNA polymerase alpha in the initiation of DNA replication. *Mol. Cell Biol.* **26**, 4843–4852 (2006).
54. Abe, T. et al. The N-terminal region of RECQL4 lacking the helicase domain is both essential and sufficient for the viability of vertebrate cells. Role of the N-terminal region of RECQL4 in cells. *Biochim. Biophys. Acta* **1813**, 473–479 (2011).
55. Castillo-Tandazo, W. et al. ATP-dependent helicase activity is dispensable for the physiological functions of Recql4. *PLoS Genet.* **15**, e1008266 (2019).
56. Sedlackova, H., Cechova, B., Mlcouskova, J. & Krejci, L. RECQ4 selectively recognizes Holliday junctions. *DNA Repair* **30**, 80–89 (2015).
57. Keller, H. et al. The intrinsically disordered amino-terminal region of human RecQL4: multiple DNA-binding domains confer annealing, strand exchange and G4 DNA binding. *Nucleic Acids Res.* **42**, 12614–12627 (2014).
58. Im, J. S. et al. Assembly of the Cdc45-Mcm2-7-GINS complex in human cells requires the Ctf4/And-1, RecQL4, and Mcm10 proteins. *Proc. Natl Acad. Sci. USA* **106**, 15628–15632 (2009).
59. Xu, X., Rochette, P. J., Feyissa, E. A., Su, T. V. & Liu, Y. MCM10 mediates RECQ4 association with MCM2-7 helicase complex during DNA replication. *EMBO J.* **28**, 3005–3014 (2009).
60. Kliszczak, M. et al. Interaction of RECQ4 and MCM10 is important for efficient DNA replication origin firing in human cells. *Oncotarget* **6**, 40464–40479 (2015).
61. Lu, H. et al. Cell cycle-dependent phosphorylation regulates RECQL4 pathway choice and ubiquitination in DNA double-strand break repair. *Nat. Commun.* **8**, 2039 (2017).
62. Lu, H. et al. RECQL4 promotes DNA end resection in repair of DNA double-strand breaks. *Cell Rep.* **16**, 161–173 (2016).
63. Shamanna, R. A. et al. RECQ helicase RECQL4 participates in non-homologous end joining and interacts with the Ku complex. *Carcinogenesis* **35**, 2415–2424 (2014).
64. De, S. et al. RECQL4 is essential for the transport of p53 to mitochondria in normal human cells in the absence of exogenous stress. *J. Cell Sci.* **125**, 2509–2522 (2012).
65. Ohlenschlager, O. et al. The N-terminus of the human RecQL4 helicase is a homeodomain-like DNA interaction motif. *Nucleic Acids Res.* **40**, 8309–8324 (2012).
66. Nguyen Ba, A. N. et al. Proteome-wide discovery of evolutionary conserved sequences in disordered regions. *Sci. Signal* **5**, rs1 (2012).
67. Davey, N. E. et al. Attributes of short linear motifs. *Mol. Biosyst.* **8**, 268–281 (2012).
68. Zarin, T. et al. Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *Elife* **8**, 1–26 (2019).
69. Uversky, V. N. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **11**, 739–756 (2002).
70. Mittag, T., Kay, L. E. & Forman-Kay, J. D. Protein dynamics and conformational disorder in molecular recognition. *J. Mol. Recognit.* **23**, 105–116 (2010).
71. Berlow, R. B., Dyson, H. J. & Wright, P. E. Functional advantages of dynamic protein disorder. *FEBS Lett.* **589**, 2433–2440 (2015).
72. Csizmok, V., Follis, A. V., Kriwacki, R. W. & Forman-Kay, J. D. Dynamic protein interaction networks and new structural paradigms in signaling. *Chem. Rev.* **116**, 6424–6462 (2016).
73. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M. & Uversky, V. N. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.* **272**, 5129–5148 (2005).
74. Fuxreiter, M. et al. Malleable machines take shape in eukaryotic transcriptional regulation. *Nat. Chem. Biol.* **4**, 728–737 (2008).
75. Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu. Rev. Biochem.* **83**, 553–584 (2014).
76. Borgia, A. et al. Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **555**, 61–66 (2018).
77. Aramburu, I. V. & Lemke, E. A. Floppy but not sloppy: Interaction mechanism of FG-nucleoporins and nuclear transport receptors. *Semin. Cell Dev. Biol.* **68**, 34–41 (2017).
78. Fung, H. Y. J., Birol, M. & Rhoades, E. IDPs in macromolecular complexes: the roles of multivalent interactions in diverse assemblies. *Curr. Opin. Struct. Biol.* **49**, 36–43 (2018).
79. Tan, P. S. et al. Two differential binding mechanisms of FG-nucleoporins and nuclear transport receptors. *Cell Rep.* **22**, 3660–3671 (2018).
80. Sottini, A. et al. Polyelectrolyte interactions enable rapid association and dissociation in high-affinity disordered protein complexes. *Nat. Commun.* **11**, 5736 (2020).
81. Uversky, V. N. Protein intrinsic disorder-based liquid-liquid phase transitions in biological systems: complex coacervates and membrane-less organelles. *Adv. Colloid Interface Sci.* **239**, 97–114 (2017).
82. Uversky, V. N. Intrinsically disordered proteins in overcrowded milieu: membrane-less organelles, phase separation, and intrinsic disorder. *Curr. Opin. Struct. Biol.* **44**, 18–30 (2017).
83. Turner, A. L. et al. Highly disordered histone H1-DNA model complexes and their condensates. *Proc. Natl Acad. Sci. USA* **115**, 11964–11969 (2018).
84. Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18**, 285–298 (2017).
85. Das, R. K., Ruff, K. M. & Pappu, R. V. Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **32**, 102–112 (2015).
86. Vuzman, D. & Levy, Y. DNA search efficiency is modulated by charge composition and distribution in the intrinsically disordered tail. *Proc. Natl Acad. Sci. USA* **107**, 21004–21009 (2010).
87. Brosh, R. M. Jr et al. Functional and physical interaction between WRN helicase and human replication protein A. *J. Biol. Chem.* **274**, 18341–18350 (1999).
88. Brosh, R. M. Jr et al. Replication protein A physically interacts with the Bloom's syndrome protein and stimulates its helicase activity. *J. Biol. Chem.* **275**, 23500–23508 (2000).
89. Doherty, K. M. et al. Physical and functional mapping of the replication protein a interaction domain of the werner and bloom syndrome helicases. *J. Biol. Chem.* **280**, 29494–29505 (2005).
90. Kang, D. et al. Interaction of replication protein A with two acidic peptides from human Bloom syndrome protein. *FEBS Lett.* **592**, 547–558 (2018).

91. Bythell-Douglas, R. & Deans, A. J. A structural guide to the Bloom syndrome complex. *Structure* **29**, 99–113 (2021).
92. Marechal, A. & Zou, L. RPA-coated single-stranded DNA as a platform for post-translational modifications in the DNA damage response. *Cell Res.* **25**, 9–23 (2015).
93. Sugitani, N. & Chazin, W. J. Characteristics and concepts of dynamic hub proteins in DNA processing machinery from studies of RPA. *Prog. Biophys. Mol. Biol.* **117**, 206–211 (2015).
94. Mer, G. et al. Structural basis for the recognition of DNA repair proteins UNG2, XPA, and RAD52 by replication factor RPA. *Cell* **103**, 449–456 (2000).
95. Ali, S. I., Shin, J. S., Bae, S. H., Kim, B. & Choi, B. S. Replication protein A 32 interacts through a similar binding interface with TIPIN, XPA, and UNG2. *Int. J. Biochem. Cell Biol.* **42**, 1210–1215 (2010).
96. Xie, S. et al. Structure of RPA32 bound to the N-terminus of SMARCAL1 redefines the binding interface between RPA32 and its interacting proteins. *FEBS J.* **281**, 3382–3396 (2014).
97. Zhao, W. et al. Promotion of BRCA2-dependent homologous recombination by DSS1 via RPA targeting and DNA mimicry. *Mol. Cell* **59**, 176–187 (2015).
98. Marino, F. et al. Structural and biochemical characterization of an RNA/DNA binding motif in the N-terminal domain of RecQ4 helicases. *Sci. Rep.* **6**, 21501 (2016).
99. Luu, K. N., Phan, A. T., Kuryavyi, V., Lacroix, L. & Patel, D. J. Structure of the human telomere in K+ solution: an intramolecular (3 + 1) G-quadruplex scaffold. *J. Am. Chem. Soc.* **128**, 9963–9970 (2006).
100. Adrian, M. et al. Structure and conformational dynamics of a stacked dimeric G-quadruplex formed by the human CEB1 minisatellite. *J. Am. Chem. Soc.* **136**, 6297–6305 (2014).
101. Kriwacki, R. W., Hengst, L., Tennant, L., Reed, S. I. & Wright, P. E. Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc. Natl Acad. Sci. USA* **93**, 11504–11509 (1996).
102. Tompa, P. & Fuxreiter, M. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.* **33**, 2–8 (2008).
103. Berjanskii, M. V. & Wishart, D. S. The RCI server: rapid and accurate calculation of protein flexibility using chemical shifts. *Nucleic Acids Res.* **35**, W531–W537 (2007).
104. Hafsa, N. E. & Wishart, D. S. CSI 2.0: a significantly improved version of the Chemical Shift Index. *J. Biomol. NMR* **60**, 131–146 (2014).
105. Shen, Y. & Bax, A. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J. Biomol. NMR* **56**, 227–241 (2013).
106. Peng, J. W. & Wagner, G. Mapping of spectral density-functions using heteronuclear NMR relaxation measurements. *J Magn Reson* **98**, 308–332 (1992).
107. Peng, J. W. & Wagner, G. Mapping of the spectral densities of N-H bond motions in eglin c using heteronuclear relaxation experiments. *Biochemistry* **31**, 8571–8586 (1992).
108. Waudby, C. A., Ramos, A., Cabrita, L. D. & Christodoulou, J. Two-dimensional NMR lineshape analysis. *Sci. Rep.* **6**, 24826 (2016).
109. Waudby, C. A. & Christodoulou, J. NMR lineshape analysis of intrinsically disordered protein interactions. *Methods Mol. Biol.* **2141**, 477–504 (2020).
110. Kim, J. Y., Meng, F., Yoo, J. & Chung, H. S. Diffusion-limited association of disordered protein by non-native electrostatic interactions. *Nat. Commun.* **9**, 4707 (2018).
111. Tzeng, S. R. & Kalodimos, C. G. Dynamic activation of an allosteric regulatory protein. *Nature* **462**, 368–372 (2009).
112. Amaral, M. et al. Protein conformational flexibility modulates kinetics and thermodynamics of drug binding. *Nat. Commun.* **8**, 2276 (2017).
113. Mendoza, O., Bourdoncle, A., Boule, J. B., Brosh, R. M. Jr. & Mergny, J. L. G-quadruplexes and helicases. *Nucleic Acids Res.* **44**, 1989–2006 (2016).
114. Do, N. Q. & Phan, A. T. Monomer-dimer equilibrium for the 5′-5′ stacking of propeller-type parallel-stranded G-quadruplexes: NMR structural study. *Chemistry* **18**, 14752–14759 (2012).
115. Bolognesi, B. et al. A concentration-dependent liquid phase separation can cause toxicity upon increased protein expression. *Cell Rep.* **16**, 222–231 (2016).
116. Sing, C. E. & Perry, S. L. Recent progress in the science of complex coacervation. *Soft Matter* **16**, 2885–2914 (2020).
117. Del Villar-Guerra, R., Trent, J. O. & Chaires, J. B. G-quadruplex secondary structure obtained from circular dichroism spectroscopy. *Angew. Chem. Int. Ed. Engl.* **57**, 7171–7175 (2018).
118. Yevdokimov, Y. M. et al. Re-entrant cholesteric phase in DNA liquid-crystalline dispersion particles. *J. Biol. Phys.* **43**, 45–68 (2017).
119. Mimura, M. et al. Quadruplex folding promotes the condensation of linker histones and DNAs via liquid-liquid phase separation. *J. Am. Chem. Soc.* **143**, 9849–9857 (2021).
120. Renaud de la Faverie, A., Guedin, A., Bedrat, A., Yatsunyk, L. A. & Mergny, J. L. Thioflavin T as a fluorescence light-up probe for G4 formation. *Nucleic Acids Res.* **42**, e65 (2014).
121. Xu, S. et al. Thioflavin T as an efficient fluorescence sensor for selective recognition of RNA G-quadruplexes. *Sci. Rep.* **6**, 24793 (2016).
122. Shen, Y. et al. Biomolecular condensates undergo a generic shear-mediated liquid-to-solid transition. *Nat. Nanotechnol.* **15**, 841–847 (2020).
123. London, T. B. et al. FANCJ is a structure-specific DNA helicase associated with the maintenance of genomic G/C tracts. *J. Biol. Chem.* **283**, 36132–36139 (2008).
124. Chang, L. W. et al. Sequence and entropy-based control of complex coacervates. *Nat. Commun.* **8**, 1273 (2017).
125. Zervoudis, N. A. & Obermeyer, A. C. The effects of protein charge patterning on complex coacervation. *Soft Matter* **17**, 6637–6645 (2021).
126. Vieregg, J. R. et al. Oligonucleotide-peptide complexes: phase control by hybridization. *J. Am. Chem. Soc.* **140**, 1632–1638 (2018).
127. Flock, T., Weatheritt, R. J., Latysheva, N. S. & Babu, M. M. Controlling entropy to tune the functions of intrinsically disordered regions. *Curr. Opin. Struct. Biol.* **26**, 62–72 (2014).
128. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29 (2015).
129. Shin, G., Jeong, D., Kim, H., Im, J. S. & Lee, J. K. RecQL4 tethering on the pre-replicative complex induces unscheduled origin activation and replication stress in human cells. *J. Biol. Chem.* **294**, 16255–16265 (2019).
130. Takai, H., Smogorzewska, A. & de Lange, T. DNA damage foci at dysfunctional telomeres. *Curr. Biol.* **13**, 1549–1556 (2003).
131. Ghosh, A. K. et al. RECQL4, the protein mutated in Rothmund-Thomson syndrome, functions in telomere maintenance. *J. Biol. Chem.* **287**, 196–209 (2012).
132. Croteau, D. L. et al. RECQL4 localizes to mitochondria and preserves mitochondrial DNA integrity. *Aging Cell* **11**, 456–466 (2012).
133. Kumari, J. et al. Mitochondrial functions of RECQL4 are required for the prevention of aerobic glycolysis-dependent cell invasion. *J. Cell Sci.* **129**, 1312–1318 (2016).

134. Petkovic, M., Dietschy, T., Freire, R., Jiao, R. & Stagljar, I. The human Rothmund-Thomson syndrome gene product, RECQL4, localizes to distinct nuclear foci that coincide with proteins involved in the maintenance of genome stability. *J. Cell Sci.* **118**, 4261–4269 (2005).
135. Singh, D. K. et al. The involvement of human RECQL4 in DNA double-strand break repair. *Aging Cell* **9**, 358–371 (2010).
136. Tsuyama, T. et al. N-terminal region of RecQ4 inhibits non-homologous end joining and chromatin association of the Ku heterodimer in Xenopus egg extracts. *Gene* **787**, 145647 (2021).
137. Falabella, M., Fernandez, R. J., Johnson, F. B. & Kaufman, B. A. Potential roles for G-quadruplexes in mitochondria. *Curr. Med. Chem.* **26**, 2918–2932 (2019).
138. Alberti, S., Gladfelter, A. & Mittag, T. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell* **176**, 419–434 (2019).
139. Ghodgaonkar, M. M. et al. Phenotypic characterization of missense polymerase-delta mutations using an inducible protein-replacement system. *Nat. Commun.* **5**, 4990 (2014).
140. Henricksen, L. A., Umbricht, C. B. & Wold, M. S. Recombinant replication protein A: expression, complex formation, and functional characterization. *J. Biol. Chem.* **269**, 11121–11132 (1994).
141. Sattler, M., Schleucher, J. & Griesinger, C. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nucl. Mag. Res. Sp.* **34**, 93–158 (1999).
142. Evangelidis, T. et al. Automated NMR resonance assignments and structure determination using a minimal set of 4D spectra. *Nat. Commun.* **9**, 384 (2018).
143. Lee, W., Tonelli, M. & Markley, J. L. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **31**, 1325–1327 (2015).
144. Korzhnev, D. M., Billeter, M., Arseniev, A. S. & Orekhov, V. Y. NMR studies of Brownian tumbling and internal motions in proteins. *Prog. Nucl. Mag. Res. Sp.* **38**, 197–266 (2001).
145. Eaton, J. W., Bateman, D., Hauberg, S. & Wehbring, R. GNU Octave version 3.8.1 manual: a high-level interactive language for numerical computations. CreateSpace Independent Publishing Platform. (2014). ISBN 1441413006, http://www.gnu.org/software/octave/doc/interpreter/.
146. Ishima, R. & Nagayama, K. Protein backbone dynamics revealed by quasi spectral density function analysis of amide N-15 nuclei. *Biochemistry* **34**, 3162–3171 (1995).
147. Ishima, R. & Nagayama, K. Quasi-spectral-density function analysis for nitrogen-15 nuclei in proteins. *J. Magn. Reson.* **108**, 73–76 (1995).
148. Farrow, N. A., Zhang, O., Szabo, A., Torchia, D. A. & Kay, L. E. Spectral density function mapping using 15N relaxation data exclusively. *J. Biomol. NMR* **6**, 153–162 (1995).
149. Ferrage, F., Cowburn, D. & Ghose, R. Accurate sampling of high-frequency motions in proteins by steady-state (15)N-{(1)H} nuclear Overhauser effect measurements in the presence of cross-correlated relaxation. *J. Am. Chem. Soc.* **131**, 6048–6049 (2009).
150. Johnson, K. A., Simpson, Z. B. & Blom, T. Global kinetic explorer: a new computer program for dynamic simulation and fitting of kinetic data. *Anal. Biochem.* **387**, 20–29 (2009).
151. Li, A., Ziehr, J. L. & Johnson, K. A. A new general method for simultaneous fitting of temperature and concentration dependence of reaction rates yields kinetic and thermodynamic parameters for HIV reverse transcriptase specificity. *J. Biol. Chem.* **292**, 6695–6702 (2017).

Acknowledgements

We thank Lukas Zidek for critical reading of the manuscript, Radovan Fiala for assistance with the NMR spectrometers and for photographing turbidity in the CD cuvette, Lukas Trantirek for helpful discussions, Klara Odehnalova for protein preparations, and Jaromír Toušek for excellent technical support. We also thank Pavel Janscak and Jean-Baptiste Boule for providing FANCJ and ScPif1 proteins, respectively. This research was supported by the Grant Agency of Masaryk University (MUNI/G/1594/2019) to K.T. and L.K. The K.T. laboratory was supported by the Czech Science Foundation (GA23-06913S). The L.K. laboratory was supported by the Czech Science Foundation (GACR 21-22593X) and Wellcome Trust Collaborative Grant (206292/E/17/Z). Z.P. was supported from the European Union's Horizon 2020 research and Innovation program TEAMING 857560, the Czech Ministry of Education, Youth and Sports TEAMING CZ.02.1.01/0.0/0.0/17_043/0009632, and Technology Agency of the Czech Republic (RETEMED TN02000122). P.K. was supported by European Regional Development Fund-Project MSCAfellow2@MUNI (No. CZ.02.2.69/0.0/0.0/18_070/0009846). C.A.W. was supported by the BBSRC (BB/T002603/1). K.K. was supported by CZ.02.01.01/00/22_008/0004575. CIISB, Instruct-CZ Centre of Instruct-ERIC EU consortium, funded by MEYS CR infrastructure project LM2023042, is gratefully acknowledged for the financial support of the measurements at the Josef Dadok National NMR Centre and at the CF Biomolecular Interactions and Crystallization.

Author contributions

A.C.P., L.K. and K.T. conceived the project and designed the experiments. A.C.P. produced most of the protein samples, performed and analyzed NMR, FA, CD, and ITC experiments. M.P. performed the CD experiments, G4 unwinding assays, Stopped-Flow measurements and EMSAs. J.C. performed the pull-down experiment. Both M.P. and J.C. performed experiments describing coacervation. R.A. generated cell lines and performed the co-immunoprecipitation experiments. C.A.W. performed the 2D lineshape NMR analysis. P.K. measured and analyzed the NMR relaxation experiments. V.M. cloned and purified the APEX2-tagged RECQ4 fragment. K.K. assisted with NMR experiments. Z.P. constructed the global kinetic model and performed numerical simulations. L.K. and K.T. supervised the research. A.C.P. and K.T. wrote the paper. All authors discussed the results and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-023-42503-z.

Correspondence and requests for materials should be addressed to Lumir Krejci or Konstantinos Tripsianes.

Peer review information *Nature Communications* thanks Takashi Nagata, Vladimir Uversky and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

Please cite this article in press as: Kumar et al., A class III ligand oscillates between internal and terminal binding modes as it engages with the Dishevelled PDZ domain, Structure (2025), https://doi.org/10.1016/j.str.2025.05.012

Structure

CellPress

Article

A class III ligand oscillates between internal and terminal binding modes as it engages with the Dishevelled PDZ domain

Jitender Kumar,[1] Miroslav Micka,[1,2] Jan Komárek,[3] Tomáš Klumpler,[1] Vojtěch Bystrý,[1] Remco Sprangers,[4] Cyril Bařinka,[3] Vítězslav Bryja,[2] and Konstantinos Tripsianes[1,5,*]

[1]CEITEC – Central European Institute of Technology, Masaryk University, Brno, Czech Republic
[2]Department of Experimental Biology, Faculty of Science, Masaryk University, Brno, Czech Republic
[3]Institute of Biotechnology of the Czech Academy of Sciences, Vestec, Czech Republic
[4]Institute of Biophysics and Physical Biochemistry, Regensburg Center for Biochemistry, University of Regensburg, Regensburg, Germany
[5]Lead contact
*Correspondence: kostas.tripsianes@ceitec.muni.cz
https://doi.org/10.1016/j.str.2025.05.012

SUMMARY

One of the largest domain-motif interactomes in human involves PSD-95/Discs-large/ZO-1 (PDZ) domains. The framework for understanding the PDZ interactome is well established; however the functional dynamics associated with PDZ-ligand interactions are poorly understood. Here, we report a dual PDZ-binding mode that ascribes unique dynamic features to class III ligand recognition. The crystal structure revealed that the PDZ domain can recognize either of the carboxylate moieties (terminal or internal) present in the class III ligand and laid out the register rules responsible for the dual recognition. Variants of the ligand designed to retain one or the other carboxylate of the native sequence were sufficient for PDZ binding. The conformational dynamics of PDZ probed by NMR relaxation dispersion experiments demonstrated that the class III ligand is shuffling binding modes as it engages with the PDZ domain. Our mechanistic findings reveal yet another aspect of PDZ binding plasticity specific to class III ligands.

INTRODUCTION

PDZ (PSD-95/Discs-large/ZO-1) domains are among the most frequent protein-protein interaction modules in the human proteome that facilitate the formation of protein networks[1–3] critical for many cellular processes,[4] including cell-cell communication,[5] cell polarity,[6,7] regulation of protein transport,[8] and protein metabolism.[9,10] The evolutionarily conserved PDZ domain typically consists of a six-stranded β barrel capped by one short and one long α helix.[11] The best-known function of PDZ domains is to organize diverse signaling complexes by recognizing a short stretch located at the extreme C terminus of target proteins.[12,13] The PDZ binding site for ligands is located within a groove formed by the long helix and one of the strands.[14] In general, the C-terminal carboxylate of a ligand is coordinated by hydrogen bonds to backbone amides of the so-called PDZ X-Φ_1-G-Φ_2 motif (X: any amino acid; Φ: hydrophobic amino acid), also termed the carboxylate-binding loop.[13,15] It is known, however, that some PDZ domains can also recognize internal sequences in their target proteins. In such cases, ligand anchoring to the PDZ X-Φ_1-G-Φ_2 motif is mediated either by a main-chain carbonyl group or a side-chain carboxylate of an aspartate or a glutamate.[16,17]

To standardize the nomenclature of PDZ interactions, the ligand residue positions are numbered from 0 at the C terminus counting backward with progressively negative integers, e.g., P_0, P_{-1}, P_{-2} etc.[18] Peptides recognized by PDZ domains usually contain a hydrophobic residue at P_0 pointing toward the adjacent pocket of the PDZ X-Φ_1-G-Φ_2 motif when the C-terminal carboxylate is engaged.[19] PDZ terminal ligands are often categorized into three major classes (denoted I, II, and III) based on the residue identity at P_{-2}.[9,15,20] Class III ligands are exceptional because they contain aspartate or glutamate at P_{-2}, thus possessing a second carboxylate in addition to the C-terminal one. It seems largely overlooked, but class III ligands offer PDZ domains the possibility to recognize one or the other carboxylate[21–23] or even exchange between the two.

PDZ domains are generally highly promiscuous modules[24] because they have a critical role in scaffolding protein complexes.[13,25] For example, the Wnt signaling protein Dishevelled (DVL) can associate with a bewildering array of partners through its PDZ domain.[26] Not only does it bind targets from all terminal classes,[27] but it also recognizes internal sequences.[28] Most importantly, DVL function is regulated by a PDZ-ligand interaction in *cis*. The DVL C-terminal tail is tucked away in its PDZ pocket,[29] acting as a DVL conformational switch that directs different biological outputs.[30] Interestingly, the DVL C-terminal tail is a class III ligand that can be reversibly capped at the C terminus by main-chain polyglutamylation.[31]

All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Please cite this article in press as: Kumar et al., A class III ligand oscillates between internal and terminal binding modes as it engages with the Dishevelled PDZ domain, Structure (2025), https://doi.org/10.1016/j.str.2025.05.012

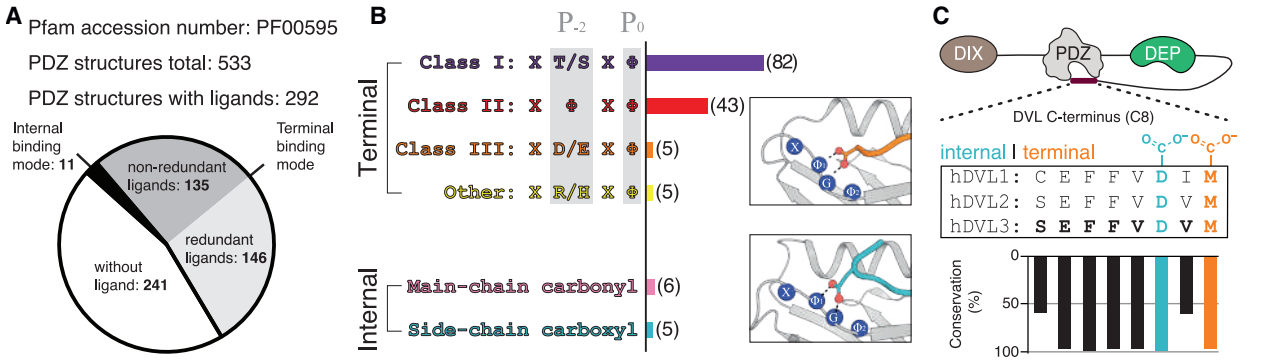


Figure 1. Binding modes of PDZ domains with cognate ligands

(A) The pie chart shows the structural information available on PDZ domains under the Pfam accession number PF00595. The structures in complex with cognate ligands report mainly on terminal binding mode (135 entries with non-redundant ligand sequences) and very few on internal binding mode (11 entries).

(B) In PDZ-ligand structures of terminal binding mode, the free C-terminal carboxylate of the ligand (P_0) is the hydrogen bond acceptor. Four distinct classes are defined for PDZ ligand interactions based on the amino acid identity at position −2 of the ligand (P_{-2}). In PDZ-ligand structures of internal binding mode, the ligand hydrogen bond acceptor is a main-chain carbonyl group of a non-terminal residue or a side-chain carboxylate of a non-terminal aspartate. Ligand recognition within PDZ domains occurs via a conserved X-Φ_1-G-Φ_2 motif (insets).

(C) The C-terminal tail of DVL is tucked away in its PDZ pocket, controlling the global conformation of the protein. The DVL C-terminal sequence contains two carboxylates (internal and terminal) as any other class III PDZ ligand. The amino acid conservation per position is shown for the last eight residues of human DVL3 protein from a multiple sequence alignment of 795 DVL sequences. The class III determinant, aspartate at P_{-2}, is fully conserved (99.4%) in the DVL protein family.

Driven by our quest to understand the multifaceted function of DVL in Wnt signaling, we set out to mechanistically characterize the DVL conformational switch. The crystal structure of the DVL3 PDZ domain in complex with the class III ligand derived from its own C terminus highlighted a dual binding mode. The PDZ domain is able to recognize either the C-terminal main-chain carboxylate or the internal side-chain carboxylate of the fully conserved aspartate at P_{-2} of the DVL3 sequence. The two binding modes are virtually indistinguishable due to two registers that fulfill highly similar binding features. Single-carboxylate variants were sufficient to anchor the ligand in the PDZ X-Φ_1-G-Φ_2 motif, as deduced by ligand-based NMR binding analysis, in support of the crystallographic observations. Relaxation dispersion experiments of PDZ-ligand fusion proteins demonstrated that the PDZ domain was in exchange between two states only when in complex with the native class III ligand. Further NMR analysis demonstrated that the class III ligand alternates between terminal and internal binding modes in an equilibrium of nearly equal populations as it engages with the PDZ domain. The present study expands the structural portrait of PDZ interactions in an unprecedented fashion, as it describes the PDZ-ligand dynamics at the binding interface, and offers interesting mechanistic perspectives about class III ligand recognition in light of the DVL function in Wnt signaling.

RESULTS

The conundrum of class III ligand recognition by PDZ domains

To better understand the PDZ binding repertoire, we analyzed the PDZ structural information available in the PDB database. A total of 533 PDB entries containing a PDZ domain were found in the Protein family database (accession number PF00595). Of these PDZ structures, 241 are in the apo-form, and 292 have a peptide ligand associated with the domain. The PDZ peptide-bound structures were manually curated (see method details) to obtain a clean dataset of non-redundant ligand sequences. In most structures (135), recognition involves the extreme C-terminal sequence of ligands, whereas recognition of internal sequences is limited to 11 structures (Figure 1A).

In the terminal binding mode, the PDZ X-Φ_1-G-Φ_2 motif wraps around the C-terminal carboxylate of a ligand.[32] For the terminal binding mode, three classes were defined earlier[9] to account for some PDZ specificity by considering the residue identity of the ligand at P_{-2} (Figures 1B and S1). The curated dataset contains 82 PDZ structures bound to class I ligands (X-S/T-X-Φ), 43 PDZ structures bound to class II ligands (X-Φ-X-Φ), and only five PDZ structures bound to class III ligands (X-D/E-X-Φ), where X denotes any amino acid and Φ is a hydrophobic amino acid (Figures 1B and S2). Thus, the dataset of PDZ structures offers limited insights into the binding mode and/or specificity of class III ligands.[33] Another five PDZ structures are bound to ligands harboring arginine or histidine at P_{-2}, which are termed "other". Overall, terminal mode recognition is too restrictive to describe the PDZ interactome[34–37] and plausible divergent binding mechanisms.

Indeed, PDZ domains show more versatility in target recognition since they are known to bind internal motifs of proteins as well.[23,28,38–40] We deduced two internal binding modes from the structural dataset based on the ligand's functional group coordinated in the X-Φ_1-G-Φ_2 motif of the PDZ domain. In six structures, the C-terminal carboxylate is substituted by a main-chain carbonyl group (Figure S3). In five structures, the side-chain carboxylate of an aspartate mimics the C-terminal carboxylate (Figure S4). The current structural portrait of the PDZ domain family suggests that in class III ligands (X-D/E-X-Φ) both main-chain C-terminal carboxylate and side-chain internal carboxylate at P_{-2} could determine PDZ binding specificity.

To understand how PDZ domains could select one or the other binding mode in class III ligands, we investigated the interaction

Please cite this article in press as: Kumar et al., A class III ligand oscillates between internal and terminal binding modes as it engages with the Dishevelled PDZ domain, Structure (2025), https://doi.org/10.1016/j.str.2025.05.012

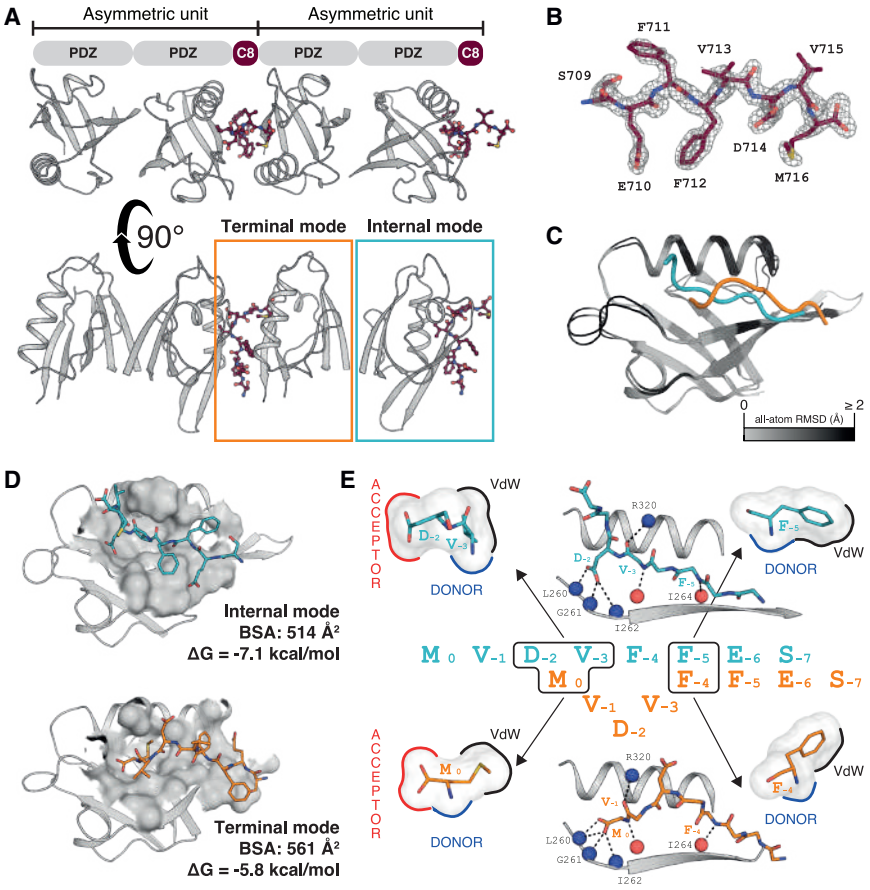


Figure 2. Class III ligand recognition by DVL3 PDZ domain in two registers

(A) Crystal structure of DVL3 PDZ domain in complex with an octapeptide corresponding to DVL3 C-terminal tail (C8). In the asymmetric unit, two molecules of the PDZ domain form a dimer, but only one is in complex with the C8 peptide (internal mode). When the crystal symmetry is expanded, the same C8 peptide is in complex with the other PDZ domain from a symmetry-related dimer (terminal mode).

(B) Electron density of C8 ligand from *Fo-Fc* map contoured at σ = 3.0 with a 1.6 Å carve radius.

(C) Superposition of the two PDZ/C8 complexes obtained from the crystallographic data (boxed in A). PDZ domains are colored with gray-to-black gradient to visualize all-atom RMSD, and the same C8 peptide is shown in cyan (internal mode) or orange (terminal mode).

(D) The C8 binding interface of each mode is shown as a gray surface on the PDZ domain. Buried surface area (BSA) and free energy (ΔG) of binding are reported.

(E) Register rules between the PDZ binding groove and C8 peptide that allow class III ligand recognition in internal (top) or terminal (bottom) mode. Dashed lines depict the hydrogen bonds between the C8 peptide and PDZ domain.

of DVL3 PDZ with its C-terminal sequence (Figure 1C). Interestingly, the class III determinant of the PDZ binding epitope, the aspartate at P_{-2}, is fully conserved across the DVL family (Figure 1C). The PDZ domain of DVL proteins is unusually promiscuous in binding terminal ligands from all classes[27,29,41] and in recognizing internal sequences due to inherent flexibility.[28] Most importantly, the DVL3 C terminus is capped reversibly by main-chain polyglutamylation,[31] providing a functional framework that may be associated with class III recognition mechanisms.

DVL3 PDZ domain exploits different registers of the class III ligand for recognition

To define the molecular interactions between the PDZ domain (aa: 243–351) and the C-terminal sequence (C8; aa: 709–716) of human DVL3, we crystallized the PDZ/C8 complex and solved the crystal structure (Figure 2). Data processing and refinement statistics are presented in Table 1. The interpretable positive *Fo-Fc* electron density peaks representing the C8 peptide were clearly visible at the 1.36 Å resolution limit. This allowed unambiguous fitting of the C8 peptide sequence in the final stages of the refinement (Figure 2B). PDZ/C8 crystals belong to the *P12₁1* space group. The asymmetric unit contains two copies of the PDZ domain forming a crystallographic dimer and a single copy of the C8 ligand bound to one of the PDZ domains (Figure 2A). Since a sizable portion of PDZ domains can form dimers,[42–46] we assessed the oligomeric state of DVL3 PDZ domain in solution (Figure S5). The crystallographic dimer was found to be inconsistent with the Small Angle X-ray Scattering (SAXS) data when comparing the overall fit to the scattering curve (χ^2 = 16). However, when the monomeric PDZ structure was considered, the structural fit to the SAXS data was dramatically improved (χ^2 = 0.9). It, therefore, appears that the DVL3 PDZ dimer is only formed under crystallization conditions and not in solution.

Within the asymmetric unit of the PDZ/C8 complex, one of the PDZ domains recognizes the internal binding motif of the ligand, where the side-chain carboxylate of aspartate at P_{-2} is accommodated in the PDZ X-Φ_1-G-Φ_2 motif. When the crystal contacts are analyzed, the C8 ligand is additionally found in a complex with the other PDZ domain from a symmetry-related molecule (Figure 2A). At the crystal interface, the main-chain C-terminal carboxylate of methionine at P_0 is recognized by the PDZ X-Φ_1-G-Φ_2 motif. Thus, the crystallographic poses may provide the molecular features of another aspect of PDZ promiscuity related to class III ligand recognition.

We, therefore, compared the internal binding mode in the asymmetric unit to the terminal binding mode via crystal operators. The two PDZ domains closely resemble each other, with a root-mean-square deviation (RMSD) of 0.65 Å over 75 C^{α} atoms (Figure 2C). Besides the loop regions of the PDZ fold, structural fluctuations are mainly observed for the side-chains of the binding groove reflecting the inherent plasticity that enables the two ligand binding modes (Figures 2C and 2D). The binding free energies for each mode were predicted from the crystallographic poses and found to differ by 1.3 kcal/mol, a value within the prediction error of the PRODIGY software.[47] Indeed, the solvent-accessible surface

Please cite this article in press as: Kumar et al., A class III ligand oscillates between internal and terminal binding modes as it engages with the Dishevelled PDZ domain, Structure (2025), https://doi.org/10.1016/j.str.2025.05.012

Table 1. X-ray refinement statistics for PDZ-C8 complex

PDB code	8S6A
	Data collection statistics
Wavelength [Å]	1.3418
Space group	*P12$_1$1*
Unit-cell parameters	a = 46.91 Å, b = 42.56 Å, c = 48.29 Å, $\alpha = \gamma = 90.0°$, $\beta = 96.1°$
Resolution limits [Å]	48.02–1.36 (1.39–1.36)
Number of unique reflections	40696 (1995)
Redundancy	7.4 (4.5)
Completeness [%]	99.9 (99.0)
CC1/2 [%]	99.8 (22.1)
$I/\sigma I$	17.3 (0.5)
R_{pim}	0.02 (1.42)
	Refinement statistics
Resolution limits [Å]	48.02–1.36 (1.40–1.36)
Total number of reflections	38661 (2973)
Number of reflections in working set	36642 (2851)
Number of reflections in test set	2019 (122)
R/R_{free} [%]	18.2/24.7 (37.7/38.9)
Total number of non-H atoms	1681
Number of non-H protein atoms	1580
Number of water molecules	101
Average B-factor [Å^2]	30.73
protein atoms	33.14
waters	36.37
Ramachandran plot [%]	
most favored	98
additionally allowed	2
disallowed	0
r.m.s. deviation bond lengths [Å]	0.014
r.m.s. deviation bond angles [°]	1.88
r.m.s. deviation planarity [Å]	0.008
r.m.s. deviation chiral centers [Å^3]	0.109
Missing residues	aa 243, 351 (chain A), aa 243–244, 269–273, 337–351 (chain B)

Values in parentheses are for the highest resolution shells.

area buried between the PDZ and the ligand is very similar in both binding modes (Figure 2D). Many of the van der Waals interactions and hydrogen bonds are common to both binding modes as well (Figure S6), in agreement with ligand binding to the PDZ groove in two different registers.

A closer look into the PDZ-ligand interactions revealed the register rules between PDZ subsites in the binding groove and ligand positions (Figure 2E). The PDZ X-Φ_1-G-Φ_2 motif forms the same hydrogen bonds with either the main-chain carboxylate and amide of M_0 (terminal mode) or the side-chain carboxylate of D_{-2} and amide of V_{-3} (internal mode). In both cases, the hydrophobic side chain of M_0 or the hydrophobic side chain of V_{-3} is buried in the adjacent PDZ pocket. The M_0 and D_{-2}/V_{-3} dipeptide adapt similar configurations matching the physicochemical properties and the spatial arrangement of their functional groups (carboxylate, amide, and methyls) (Figure 2E). The carbonyls of V_{-1} (terminal mode) and V_{-3} (internal mode) coincide and form another hydrogen bond with a guanidino group projecting from the long PDZ helix. After this point, the main-chain of the ligand follows a different trajectory in each register, possibly due to torsional restraints imposed once the carboxylate (main-chain or side chain) is anchored to the PDZ X-Φ_1-G-Φ_2 motif. In the internal binding mode, the C8 ligand exhibits an extended conformation as seen for most PDZ-ligand interactions, whereas, in the terminal binding mode, the C8 ligand bends around D_{-2}.[33] The ligand chains in each binding mode are offset by one residue with respect to each other as they align with the strand of the PDZ groove. The spatial positions of F_{-4} (terminal mode) and F_{-5} (internal mode) match well to have in common another hydrogen bond and ring contacts with the PDZ groove (Figure 2E).

Each binding mode, terminal or internal, is sufficient for PDZ recognition

The crystallographic analysis does not clarify whether the observed binding modes result from crystal packing effects or also occur in solution, where they may have functional significance. We therefore designed variants of the class III ligand devoid of the terminal, the internal, or both carboxylates, and compared their PDZ binding by NMR to the putative dual binding mode of the wild-type ligand. For ligand-observed NMR titration experiments, we generated longer peptide forms that could be isotopically labeled and purified to homogeneity. These involved the last 24 amino acids of the native DVL3 C terminus (C24; dual mode), a C-terminal main-chain polyglutamylated variant of biological relevance[31] that shifts the terminal carboxylate away from the native sequence (C24-12E; internal mode), a single point mutant that removes the side-chain carboxylate (C24 D714T; terminal mode), and their combination as control of carboxylate-mediated binding (C24-12E D714T; none of the modes). Overlay of the ^{1}H-^{15}N Heteronuclear Single Quantum Correlation (HSQC) spectra of the four ligand versions showed distinct chemical shifts changes adjacent to the polyglutamylated extension or around the single-point mutation (Figure S7). Overall, there is a narrow range of peak dispersion in the ^{1}H-^{15}N HSQC spectra, suggesting that all four peptides are intrinsically disordered (Figures 3A and S7).

Since the concentration of ligand peptides could not be determined by optical absorption measurements at a wavelength of 280 nm, to be as accurate as possible, their concentrations were referenced and matched to a synthesized C24 peptide by comparing their NMR signals (Figures S7A and S7B). We then recorded a series of ^{1}H-^{15}N HSQC or ^{1}H-^{13}C HSQC spectra with 50 μM of each peptide (^{15}N, ^{15}N and ^{13}C-methyl-LV, or ^{15}N and ^{13}C-methyl-M labeled) with increasing amounts of unlabeled DVL3 PDZ domain. In common, the titrations of ligands bearing both or any of the carboxylates were characterized by peak broadening accompanied by small chemical shift perturbations (Figure 3), reporting directly on the interaction with the PDZ domain. In C24 (dual), C24 D714T (terminal), and C24-12E

Please cite this article in press as: Kumar et al., A class III ligand oscillates between internal and terminal binding modes as it engages with the Dishevelled PDZ domain, Structure (2025), https://doi.org/10.1016/j.str.2025.05.012

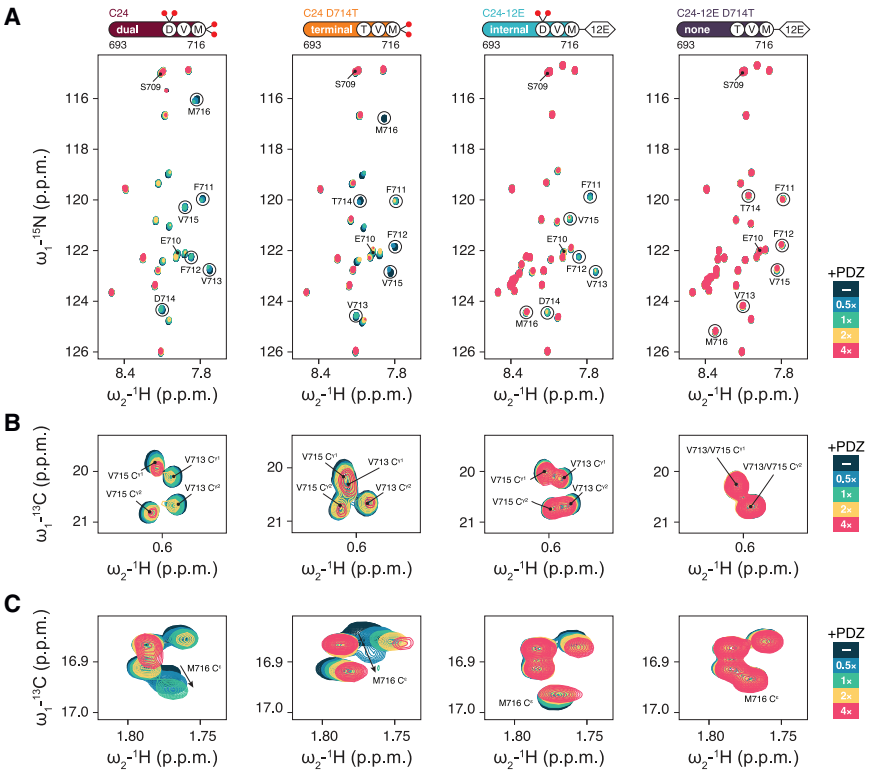


Figure 3. PDZ binding to ligand variants monitored by NMR

50 μM of labeled peptides corresponding to the last 24 residues of DVL3 sequence, wild-type (C24) or variants (C24 D714T, C24-12E, and C24-12E D714T), were serially titrated with increasing amounts of DVL3 PDZ domain monitoring the backbone amide groups (A), the methyl groups of valines (B), or the methyl group of methionines (C). Peptide cross-peaks corresponding to the last eight residues of the native sequence that were complexed with PDZ in the crystal structure are labeled in all spectra (methyl numbers of valines are random, they do not refer to stereospecific assignments).

(internal) peptides, peak intensity decreased rapidly for the residues that are part of the PDZ binding epitope seen in the two crystallographic poses (Figure 3). In addition, broadening was observed for the cross-peaks of residues preceding the PDZ-binding epitope (Figure S8). These secondary effects could be ascribed to non-specific ligand interactions. Last, the C24-12E D714T peptide version devoid of both carboxylates showed small changes in peak intensity during the PDZ titration. Taken together, the NMR titration data reinforce the notion that the carboxylates are the principal determinants of the interaction, yet each of them (terminal or internal) is sufficient for class III ligand recognition by the DVL3 PDZ domain.

To obtain a more qualitative picture of the binding mode related to each or both carboxylates, we analyzed the peak intensity of peptide backbone amides and methyls as a function of the PDZ molar ratio in the course of the NMR titration (Figure S9). For the last eight amino acids of the native sequence, ligand peak intensity changes (I/I_0; where I and I_0 are intensities in the presence and absence of added PDZ domain, respectively) report on the direct contacts with the PDZ domain observed in the crystallographic poses. The comparative binding analysis offers several interesting findings when coupled to the crystallographic data (Figure S9). First, the backbone amide and methyl of M716 (last amino acid of the native sequence) shows a steep intensity decrease when PDZ is titrated to C24 or C24 D714T peptides but a shallow linear effect when added to their polyglutamylated counterparts. The two distinct responses of M716 are in support of PDZ terminal binding mode for C24 and C24 D714T peptides but not for the polyglutamylated versions, confirming the design of variants. Second, the C24-12E showed sharper decrease of peak intensities from D714 backward, characteristic of the internal binding, while C24-12E D714T displayed linear effects throughout the binding epitope indicating minimal interaction. Third, the peak intensity trends converged from V713 backward for the variants bearing any or both the carboxylates, as the two registers observed in crystallographic analysis coincide around this point with an offset of one residue and sense the same PDZ environment. Finally, yet importantly, C24 (dual) consistently lagged behind C24 D714T (terminal) in the peak intensity decrease of the last three residues (M716, V715, and D714). This lag may represent the weighted population average of terminal and internal binding modes as C24 peptide (dual) alternates between the bound conformations related to the two registers.

The spectroscopic behavior supports different binding modes among the ligands; however, peak broadening (decrease of peak intensity) is also sensitive to the magnitude of the chemical shift difference (between free and bound peptide forms) or other exchange processes, such as on/off rates related to register shifts of C24 (dual). In addition, a statistical analysis of the peak broadening profiles indicated that the observed differences between C24 (dual) and C24 D714T (terminal) are not significant, although the small sample size (n = 3) reduces the likelihood of finding a significant p-value (Figure S9B). With these limitations and considerations in mind, we sought additional means to discern the putative dual binding mode between the PDZ domain and the wild-type ligand.

Class III ligand shuffles binding modes as it engages with the PDZ domain

The NMR titration could not be analyzed quantitatively because the ligand peaks experience severe line broadening, likely caused by the exchange between the free and bound ligand states. In order to decipher the dynamics of the dual binding mode, the exchange contributions from binding-unbinding of the ligand must be suppressed by studying the system on saturation. To this end, the ligands (C24, dual mode; C24 D714T, terminal mode; and C24-12E, internal mode) were fused at the C terminus of the DVL3 PDZ domain. This is a common strategy in obtaining PDZ-ligand structures[28,33,45,48,49] because covalent

Please cite this article in press as: Kumar et al., A class III ligand oscillates between internal and terminal binding modes as it engages with the Dishevelled PDZ domain, Structure (2025), https://doi.org/10.1016/j.str.2025.05.012

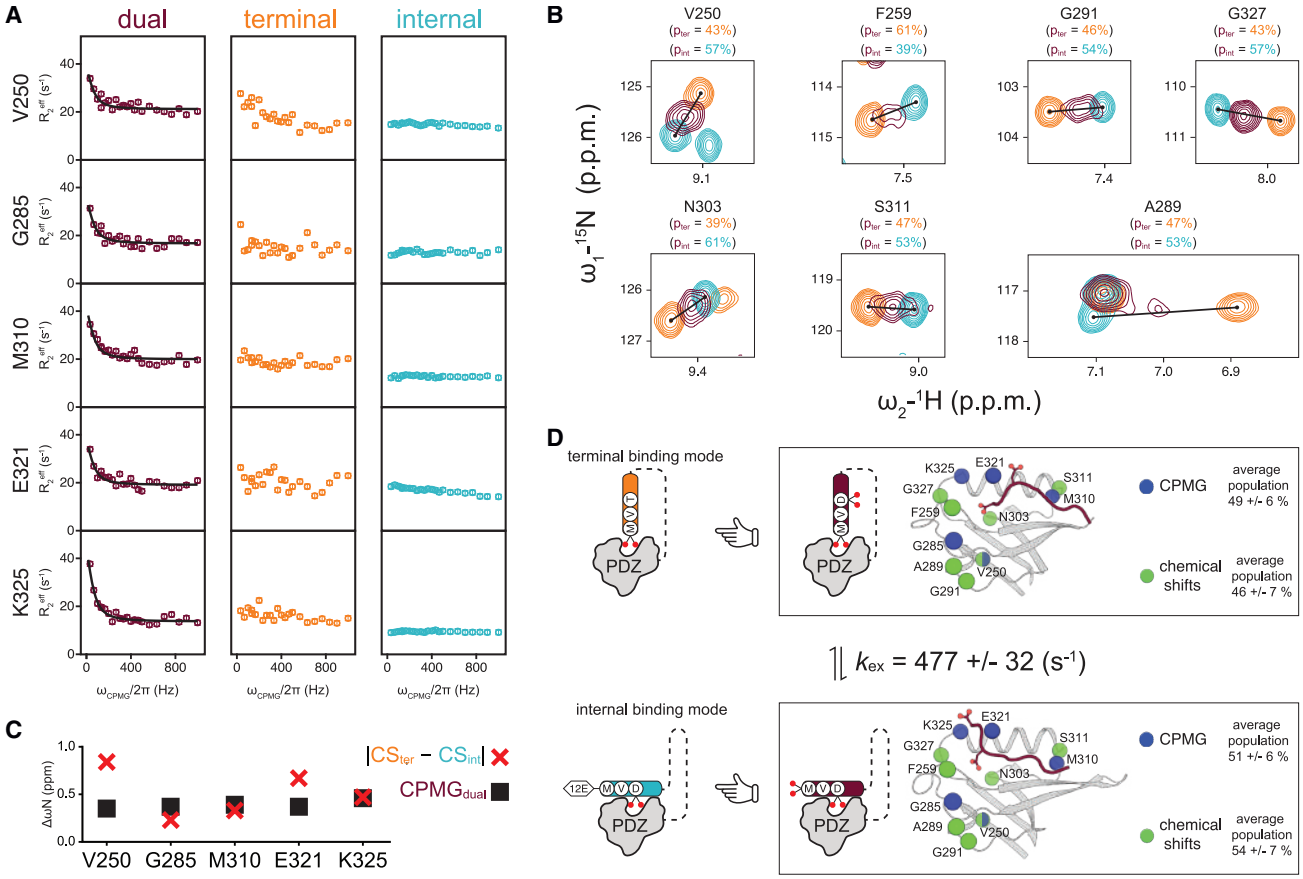


Figure 4. The class III ligand alternates between terminal and internal modes as it binds to the DVL3 PDZ domain

(A) ^{15}N CPMG relaxation dispersion data were measured for PDZ+C24, PDZ+C24 D714T, and PDZ+C24-12E. The dispersion profiles of five PDZ residues are shown. Error bars denote the uncertainty of the measured $R_{2,eff}$ value at each CPMG frequency.

(B) Selected PDZ peaks that are perturbed because of ligand binding in *cis*. The PDZ+C24 peaks are in exchange between terminal and internal binding modes of the ligand. The distance in the linear path from PDZ+C24 D714T (terminal binding) or PDZ+C24-12E (internal binding) peak positions reports on the population of each binding mode in PDZ+C24.

(C) ^{15}N chemical shift differences between terminal (PDZ+C24 D714T) and internal (PDZ+C24-12E) binding modes and the ones fitted from the CPMG data of the dual binding mode (PDZ+C24) are plotted for five PDZ residues.

(D) Summary of the dual recognition mode between the PDZ domain and the class III ligand based on crystallographic data, binding analysis of ligand variants, and NMR relaxation dispersion data.

coupling of the two interacting moieties shifts the equilibrium to the bound state, and stoichiometric complexes are formed even for low affinity interactions.[50] Analysis by native mass spectrometry demonstrated that similar to the PDZ domain, the oligomeric state of the PDZ fusion proteins was predominantly monomeric (Figure S10). The fingerprint spectrum of the PDZ domain fused to the native ligand (PDZ+C24) was distinct from that of the free PDZ domain or the free ligand in support of intramolecular binding (Figure S11).

Next, the nature of the PDZ-C24 interaction in *cis* was compared to the binding in *trans* by performing a PDZ titration with unlabeled C8 ligand (Figure S12). The PDZ peaks in PDZ+C24 spectra were part of the linear binding trajectory observed in the titrations with the C8 ligand. Hence, the interaction was faithfully reproduced at stoichiometry in *cis*, but their chemical shift change was always larger than the last point of the titration (Figures S12A and S12B). The observed chemical shift changes in PDZ+C24 indicated that saturation of binding is attainable in *cis*, whereas in *trans* it reaches at most 75% (Figure S12C). To saturate PDZ binding in *trans* and isolate exchange contributions from register shift would require a very high concentration of the C8 ligand, which was not practically feasible. Apart from PDZ+C24 (dual), binding in *cis* was also confirmed for PDZ+C24 D714T (terminal) and PDZ+C24-12E (internal). Interestingly, the PDZ chemical shifts observed among the fused variants were different suggesting a dynamic PDZ interface accommodating alternative binding modes (Figure S11).

Next, nitrogen Carr-Purcell-Meiboom-Gill (CPMG) relaxation dispersion experiments were recorded to interpret line broadening in the NMR spectra purely in light of the different binding modes of the ligands. For the PDZ+C24 D714T and PDZ+C24-12E fused variants, where a single binding mode is expected, the ^{15}N CPMG relaxation profiles were flat, consistent with a stable complex that is not dynamic on the millisecond timescale. In

Please cite this article in press as: Kumar et al., A class III ligand oscillates between internal and terminal binding modes as it engages with the Dishevelled PDZ domain, Structure (2025), https://doi.org/10.1016/j.str.2025.05.012

contrast, a number of CPMG relaxation dispersion profiles was observed for the PDZ+C24 fused protein (Figure 4A). Five dispersion curves were analyzed using a two-site exchange model, where the C24 ligand alternates between internal and terminal binding modes, as suggested by the crystal structure. This revealed that one mode is populated to 49 ± 6%, the other mode is populated to 51 ± 6%, and that the exchange rate (k_{ex} = $k_{internal \rightarrow terminal}$ + $k_{terminal \rightarrow internal}$) between the two binding modes is 477 ± 32 s^{-1} (Figure S13). Thus, the class III ligand dynamically exchanges between two binding modes, as it interacts with the PDZ domain.

To independently confirm these results and obtain structural information about the alternate binding modes, we compared the different chemical shifts induced in the PDZ domain when fused to eachof the ligands (PDZ+C24, PDZ+C24 D714T, or PDZ+C24-12E) (Figure S14). In most instances, PDZ+C24 peaks laid along linear paths between the peak positions of PDZ+C24 D714T and PDZ+C24-12E variants (Figure 4B). This phenomenon is characteristic of two distinct states of a protein in exchange between two binding partners.[51] In the fused proteins, PDZ chemical shifts are exquisitely sensitive, not only to the different binding modes (terminal or internal), but also to the local electronic environment of the ligand as it alternates between them. Each fused variant (PDZ+C24 D714T and PDZ+C24-12E) represents one of the binding modes corresponding to the crystallographic poses as determined by the carboxylate moiety (internal or terminal) present in the ligand sequence. PDZ domain fused to class III ligand (PDZ+C24) is in exchange between the two binding modes and the populations of the interconverting species are directly inferred from the chemical shift differences relative to the internal or terminal modes observed in the NMR spectra (Figure 4B). The ^{15}N chemical shift differences between terminal (PDZ+C24 D714T) and internal (PDZ+C24-12E) binding modes obtained from the spectra and the fitted values derived from the CPMG data of the dual binding mode (PDZ+C24) are compared in Figure 4C. Chemical shifts and their linear relationship along with the parameters extracted from the CPMG data, support that class III ligand recognition by DVL3 PDZ domain alternates between the two binding modes in an equilibrium of nearly equal populations (Figure 4D).

DISCUSSION

Early attempts to deduce a selectivity pattern in PDZ interactions led to a classification system based on the hydrophobic nature of the terminal residue (P_0) and the identity of the P_{-2} residue of ligands.[12,15,33] However, this classification is insufficient to describe the interactomic space that comprises 266 PDZ domains and about 4000 putative terminal ligands.[1] The PDZ interactome is highly promiscuous as one PDZ domain may bind several C-terminal sequences, while different PDZ domains may also recognize the same ligand. Yet, the space of PDZ interactions is even larger, including binding to internal sequences of proteins[4,17,28] or lipids,[52,53] such as phosphatidylinositol and cholesterol, and it is also amenable to regulation by phosphorylation[36,54,55] or allosteric interactions.[56–58] Regardless of the overall complexity, the simplified classification captures an important chemical feature unique to class III ligands: the presence of a second carboxylate moiety adjacent to the C-terminal one. Although pertinent to PDZ-ligand interactions, this notion has not been discussed as a concept in PDZ domain structure, specificity, and cellular function. To put it simply, given a class III ligand sequence, is PDZ recognition mediated by the terminal or the internal carboxylate moiety?

The present study on the intramolecular interaction of DVL3 protein offers fresh insights into the documented plasticity of PDZ-ligand interactions.[36,54,59] The class III ligand exchanges between terminal and internal binding modes as it engages with the PDZ domain in an equilibrium of nearly equal populations. Dual ligand recognition is explained by two binding registers that form highly similar interactions when each carboxylate is anchored to the PDZ X-Φ_1-G-Φ_2 motif. As long as the binding energetics of the registers do not differ significantly (Figure 2D), three factors may explain the continuous shuffling of the class III ligand within the PDZ groove.

First, the proximity of the two carboxylate moieties. At any given time, one carboxylate is anchored to the PDZ X-Φ_1-G-Φ_2 motif, but the other is always in the vicinity of this motif (Figure 2E). Under equilibrium conditions, as one ligand register dissociates from the PDZ domain, the second register takes over PDZ binding by anchoring the available carboxylate and vice versa. Since the carboxylates within the same ligand molecule interchange positions in the PDZ X-Φ_1-G-Φ_2 motif (Figure 4), the binding kinetics are no longer governed by molecular diffusion only. Instead, they likely reflect external cooperativity effects induced by the increased local concentration of the carboxylate moiety.[60] If the carboxylates are placed further apart in the sequence, the effect is expected to attenuate even if two binding registers are still supported. Therefore, this phenomenon should be common and specific to class III ligands by virtue of their classification.

Second, the intrinsic dynamics of the PDZ domain. It has been suggested that the conformational flexibility of the DVL2 PDZ domain allows it to accommodate either terminal or internal ligands.[61] Indeed, when the DVL3 binding modes are compared, the main differences in PDZ involve the side-chain conformations of residues in the binding groove (Figures 2C, 2D, and S6). These observations suggest that the intrinsic dynamics of DVL3 and other PDZ domains,[61–63] allow to sample conformational states required to accommodate the class III ligand in different registers or to support promiscuous ligand binding in general.

Third, the nature of the ligand sequence. Between the modes, terminal methionine, and internal aspartate along with the preceding valine, adopt similar configurations as they bind to the PDZ domain. The length and linear form of the side chain give methionine a great deal of torsional freedom to orient itself in the adjacent PDZ pocket as the internal valine and to engage at the same time with the PDZ X-Φ_1-G-Φ_2 motif as the internal aspartate (Figure 2E). However, methionine is rarely observed at the C terminus of sequences recognized by PDZ domains (Figure S1).[37] The prevalence of isoleucine, leucine, or valine at the C terminus may require the less bulky but still hydrophobic alanine preceding the internal aspartate to fulfill the register rules. Similar considerations should apply if the internal carboxylate is part of the longer glutamate. Another important feature in the DVL3 ligand sequence seems to be the two consecutive phenylalanines. They coincide as the two registers align with the

Please cite this article in press as: Kumar et al., A class III ligand oscillates between internal and terminal binding modes as it engages with the Dishevelled PDZ domain, Structure (2025), https://doi.org/10.1016/j.str.2025.05.012

strand of the PDZ groove with an offset of one residue. It might be that they balance the binding energetics between the registers. It should, however, be noted that six out of the last eight amino acids are highly conserved (97% or higher) in the three DVL proteins (Figure 1C). The conservation may imply sequence attributes in relation to binding registers, but it also strongly suggests a functional role for the dual binding mode in DVL biology.

DVL proteins are key components of Wnt signaling pathways which relay the signal downstream of membrane receptors.[26,64,65] It is widely accepted that Wnt-dependent activation of DVLs involves global conformational rearrangements,[30,66,67] which could be regulated by post-translational modifications.[30,68,69] The intramolecular interaction studied here, between the PDZ domain and its C terminus, keeps DVLs in an autoinhibited state,[29] presumably by blocking PDZ binding to target proteins (Figure 1C). The so-called "closed" conformation of DVLs is less active in both Wnt/β-catenin and Wnt/planar cell polarity pathways,[29,70] and is associated with the increased capacity of DVLs to form biological condensates.[30] Phosphorylation of DVLs in both the PDZ domain and the conserved C terminus[30,31,71] affects the DVL conformational dynamics triggering an "open" conformation. It therefore appears that the DVL conformational switch modulates the activation of both the Wnt/β-catenin and Wnt/planar cell polarity pathways in response to phosphorylation.

Mechanistically, the dual binding mode described here may allow the selective tuning of the DVL function by another post-translational modification, that of C-terminal main-chain polyglutamylation. It has been recently shown that the TTLL11 enzyme extends the primary sequence of DVL3 by the addition of glutamates ranging in numbers from one to more than ten.[31] This modification modulates the phase separation and phosphorylation of DVL3, two effects reversed by the action of the deglutamylating enzyme CCP6. At the cellular and/or animal level, a polyglutamylation-mimicking DVL3 with 12 C-terminally added glutamates (DVL3-12E) affected the Wnt/planar cell polarity pathway but not the Wnt/β-catenin pathway.[31] Our *in vitro* data show that the DVL3 polyglutamylated peptide variant supports binding to the PDZ domain by utilizing the internal binding mode (Figures 3, S8, and S9). Thus, the underlying mechanism of DVL activation does not need to be different when the C terminus is polyglutamylated since the DVL conformational switch could operate via the internal binding mode. Interestingly, a DVL3 double mutant E710A/D714A, which presumably restricts the intramolecular interaction to the terminal mode, could not be polyglutamylated and phenotypically resembled the "closed" DVL conformation.[31] Polyglutamylation coupled to internal binding may additionally exert a combinatory effect on the conformational switch since it augments phosphorylation of the DVL C terminus by casein kinase 1ε and prevents DVL from undergoing phase separation.[31]

Inappropriate activation of Wnt signaling by upregulation of DVLs has been implicated in tumorigenesis and cancer.[72–76] Therefore, there is great therapeutic value in targeting DVLs and their PDZ domain-mediated interactions.[10] The present study could impact such efforts because it offers insights into class III ligand recognition in two different registers by means of PDZ conformational dynamics and common principles associated with molecular recognition (Figure 2). Understanding the dynamical and molecular determinants of PDZ interactions is crucial in the modern design of drug-like molecules.

Limitations of the study

The crystal structure of the DVL3 PDZ domain in complex with a class III ligand suggested the possibility of a dual binding mode. This has been attributed to a two-state exchange process because the peak positions in the HSQC spectra of the wild-type (dual mode) were intermediate of the mutants representing either the internal or terminal mode. In addition, only the wild-type produced relaxation dispersion profiles, indicating a dynamic exchange between binding modes, whereas the two mutants did not. However, the differences in ^{15}N chemical shifts fitted from the CPMG data of the wild-type are not in full agreement with the observed differences in the spectra of the mutants (Figure 4C). This implies that the two-state exchange model might be a simplification of the binding dynamics. A more elaborate analysis would require additional information, e.g., kinetics or thermodynamics, in combination with the dual binding mechanism.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Konstantinos Tripsianes (kostas.tripsianes@ceitec.muni.cz).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Atomic coordinates (PDB ID: 8S6A) and SAXS data collection (SASBDB accession code: SASDWG9) are publicly available.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank Ondrej Šedo, David Potěšil, and Zbyněk Zdráhal for native MS measurements and analysis. This research was supported by Czech Science Foundation grants GA23-06913S (K.T.) and GA22-25365S (C.B. or V. Bryja), Czech Academy of Sciences grant RVO: 86652036 (C.B.). J.Komárek was supported by the Czech Academy of Sciences (Support of Perspective Human Resources Program). K.T. acknowledges financial support from EU HORIZON-WIDERA-2023-ACCESS-04-01 - Pathways to Synergies "MILESTONE" program (101159708). CIISB, Instruct-CZ Centre of Instruct-ERIC EU consortium, funded by MEYS CR infrastructure project LM2023042, and European Regional Development Fund-Project "Innovation of Czech Infrastructure for Integrative Structural Biology" (no. CZ.02.01.01/00/23_015/0008175) is gratefully acknowledged for the financial support of the measurements at Josef Dadok National NMR Center (CEITEC), CF Biomolecular Interactions and Crystallography (CEITEC), CF Proteomics (CEITEC), and CMS-BIOCEV ("Biophysical techniques, Crystallization, Diffraction, Structural mass spectrometry"). NCMG research infrastructure project LM2023067 for data analysis by CF Bioinformatics (CEITEC).

AUTHOR CONTRIBUTIONS

J. Kumar, V. Bryja, and K.T. conceived and designed the project. J. Kumar curated the PDB database, produced protein samples, performed, and analyzed all NMR experiments. M.M. performed NMR chemical shift assignments of the ligands. J. Komárek and C.B. setup crystallization and

Please cite this article in press as: Kumar et al., A class III ligand oscillates between internal and terminal binding modes as it engages with the Dishevelled PDZ domain, Structure (2025), https://doi.org/10.1016/j.str.2025.05.012

determined the X-ray structure. T.K. performed SAXS experiments. V. Bystrý performed sequence conservation analysis. R.S. performed and analyzed CPMG experiments. R.S., C.B., V. Bryja, and K.T. supervised research. J. Kumar. and K.T. wrote the paper with contributions from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Curation and classification of PDZ-containing structures deposited in the PDB database
  - Conservation analysis of DVL C-terminus
  - Molecular cloning
  - Protein expression and purification
  - Crystallization and data collection
  - Structure solution and refinement
  - Small angle X-Ray scattering
  - Native mass spectroscopy
  - NMR spectroscopy
  - CPMG relaxation dispersion experiments
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.str.2025.05.012.

Received: April 3, 2024
Revised: March 5, 2025
Accepted: May 20, 2025

REFERENCES

1. Gogl, G., Zambo, B., Kostmann, C., Cousido-Siah, A., Morlet, B., Durbesson, F., Negroni, L., Eberling, P., Jané, P., Nominé, Y., et al. (2022). Quantitative fragmentomics allow affinity mapping of interactomes. Nat. Commun. *13*, 5472. https://doi.org/10.1038/s41467-022-33018-0.
2. Harrison, S.C. (1996). Peptide-surface association: the case of PDZ and PTB domains. Cell *86*, 341–343. https://doi.org/10.1016/s0092-8674(00)80105-1.
3. Kennedy, M.B. (1995). Origin of PDZ (DHR, GLGF) domains. Trends Biochem. Sci. *20*, 350. https://doi.org/10.1016/s0968-0004(00)89074-x.
4. Lee, H.J., and Zheng, J.J. (2010). PDZ domains and their binding partners: structure, specificity, and modification. Cell Commun. Signal. *8*, 8. https://doi.org/10.1186/1478-811X-8-8.
5. Kim, E., and Sheng, M. (2004). PDZ domain proteins of synapses. Nat. Rev. Neurosci. *5*, 771–781. https://doi.org/10.1038/nrn1517.
6. Wu, H., Feng, W., Chen, J., Chan, L.N., Huang, S., and Zhang, M. (2007). PDZ domains of Par-3 as potential phosphoinositide signaling integrators. Mol. Cell *28*, 886–898. https://doi.org/10.1016/j.molcel.2007.10.028.
7. Bilder, D. (2001). PDZ proteins and polarity: functions from the fly. Trends Genet. *17*, 511–519. https://doi.org/10.1016/s0168-9525(01)02407-6.
8. Fernandez-Larrea, J., Merlos-Suarez, A., Urena, J.M., Baselga, J., and Arribas, J. (1999). A role for a PDZ protein in the early secretory pathway for the targeting of proTGF-alpha to the cell surface. Mol Cell *3*, 423–433. https://doi.org/10.1016/s1097-2765(00)80470-0.
9. Nourry, C., Grant, S.G.N., and Borg, J.P. (2003). PDZ domain proteins: plug and play. Sci. STKE *2003*, RE7. https://doi.org/10.1126/stke.2003.179.re7.
10. Christensen, N.R., Čalyševa, J., Fernandes, E.F.A., Lüchow, S., Clemmensen, L.S., Haugaard-Kedström, L.M., and Strømgaard, K. (2019). PDZ Domains as Drug Targets. Adv. Ther. *2*, 1800143. https://doi.org/10.1002/adtp.201800143.
11. Ponting, C.P. (1997). Evidence for PDZ domains in bacteria, yeast, and plants. Protein Sci. *6*, 464–468. https://doi.org/10.1002/pro.5560060225.
12. Ranganathan, R., and Ross, E.M. (1997). PDZ domain proteins: scaffolds for signaling complexes. Curr. Biol. *7*, R770–R773. https://doi.org/10.1016/s0960-9822(06)00401-5.
13. Harris, B.Z., and Lim, W.A. (2001). Mechanism and role of PDZ domains in signaling complex assembly. J. Cell Sci. *114*, 3219–3231. https://doi.org/10.1242/jcs.114.18.3219.
14. Gutierrez-Gonzalez, L.H., Rivas-Fuentes, S., Guzman-Beltran, S., Flores-Flores, A., Rosas-Garcia, J., and Santos-Mendoza, T. (2021). Peptide Targeting of PDZ-Dependent Interactions as Pharmacological Intervention in Immune-Related Diseases. Molecules *26*, 6367. https://doi.org/10.3390/molecules26216367.
15. Songyang, Z., Fanning, A.S., Fu, C., Xu, J., Marfatia, S.M., Chishti, A.H., Crompton, A., Chan, A.C., Anderson, J.M., and Cantley, L.C. (1997). Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. Science *275*, 73–77. https://doi.org/10.1126/science.275.5296.73.
16. Banerjee, M., Zoetewey, D.L., Ovee, M., Mazumder, S., Petrenko, V.A., Samoylova, T.I., and Mohanty, S. (2012). Specificity and promiscuity in human glutaminase interacting protein recognition: insight from the binding of the internal and C-terminal motif. Biochemistry *51*, 6950–6960. https://doi.org/10.1021/bi3008033.
17. Ali, M., McAuley, M.M., Lüchow, S., Knapp, S., Joerger, A.C., and Ivarsson, Y. (2021). Integrated analysis of Shank1 PDZ interactions with C-terminal and internal binding motifs. Curr. Res. Struct. Biol. *3*, 41–50. https://doi.org/10.1016/j.crstbi.2021.01.001.
18. Aasland, R., Abrams, C., Ampe, C., Ball, L.J., Bedford, M.T., Cesareni, G., Gimona, M., Hurley, J.H., Jarchau, T., Lehto, V.P., et al. (2002). Normalization of nomenclature for peptide motifs as ligands of modular protein domains. FEBS Lett. *513*, 141–144. https://doi.org/10.1016/s0014-5793(01)03295-1.
19. Reina, J., Lacroix, E., Hobson, S.D., Fernandez-Ballester, G., Rybin, V., Schwab, M.S., Serrano, L., and Gonzalez, C. (2002). Computer-aided design of a PDZ domain to recognize new target sequences. Nat. Struct. Biol. *9*, 621–627. https://doi.org/10.1038/nsb815.
20. Hung, A.Y., and Sheng, M. (2002). PDZ domains: structural modules for protein complex assembly. J. Biol. Chem. *277*, 5699–5702. https://doi.org/10.1074/jbc.R100065200.
21. Umnajvijit, W., Sangthong, J., Loison, F., Vaeteewoottacharn, K., and Ponglikitmongkol, M. (2021). An internal class III PDZ binding motif in HPV16 E6* protein is required for Dlg degradation activity. Biochim. Biophys. Acta. Gen. Subj. *1865*, 129850. https://doi.org/10.1016/j.bbagen.2021.129850.
22. Lemaire, J.F., and McPherson, P.S. (2006). Binding of Vac14 to neuronal nitric oxide synthase: Characterisation of a new internal PDZ-recognition motif. FEBS Lett. *580*, 6948–6954. https://doi.org/10.1016/j.febslet.2006.11.061.
23. Mu, Y., Cai, P., Hu, S., Ma, S., and Gao, Y. (2014). Characterization of diverse internal binding specificities of PDZ domains by yeast two-hybrid screening of a special peptide library. PLoS One *9*, e88286. https://doi.org/10.1371/journal.pone.0088286.
24. Basu, M.K., Carmel, L., Rogozin, I.B., and Koonin, E.V. (2008). Evolution of protein domain promiscuity in eukaryotes. Genome Res. *18*, 449–461. https://doi.org/10.1101/gr.6943508.
25. Schreiber, G., and Keating, A.E. (2011). Protein binding specificity versus promiscuity. Curr. Opin. Struct. Biol. *21*, 50–61. https://doi.org/10.1016/j.sbi.2010.10.002.

Please cite this article in press as: Kumar et al., A class III ligand oscillates between internal and terminal binding modes as it engages with the Dishevelled PDZ domain, Structure (2025), https://doi.org/10.1016/j.str.2025.05.012

26. Sharma, M., Castro-Piedras, I., Simmons, G.E., Jr., and Pruitt, K. (2018). Dishevelled: A masterful conductor of complex Wnt signals. Cell. Signal. *47*, 52–64. https://doi.org/10.1016/j.cellsig.2018.03.004.

27. Mieszczanek, J., Strutt, H., Rutherford, T.J., Strutt, D., Bienz, M., and Gammons, M.V. (2022). Selective function of the PDZ domain of Dishevelled in noncanonical Wnt signalling. J. Cell Sci. *135*, jcs259547. https://doi.org/10.1242/jcs.259547.

28. Zhang, Y., Appleton, B.A., Wiesmann, C., Lau, T., Costa, M., Hannoush, R. N., and Sidhu, S.S. (2009). Inhibition of Wnt signaling by Dishevelled PDZ peptides. Nat. Chem. Biol. *5*, 217–219. https://doi.org/10.1038/nchembio.152.

29. Lee, H.J., Shi, D.L., and Zheng, J.J. (2015). Conformational change of Dishevelled plays a key regulatory role in the Wnt signaling pathways. eLife *4*, e08142. https://doi.org/10.7554/eLife.08142.

30. Harnos, J., Canizal, M.C.A., Jurasek, M., Kumar, J., Holler, C., Schambony, A., Hanakova, K., Bernatik, O., Zdrahal, Z., Gomoryova, K., et al. (2019). Dishevelled-3 conformation dynamics analyzed by FRET-based biosensors reveals a key role of casein kinase 1. Nat. Commun. *10*, 1804. https://doi.org/10.1038/s41467-019-09651-7.

31. Kravec, M., Šedo, O., Nedvědová, J., Micka, M., Šulcová, M., Zezula, N., Gömöryová, K., Potěšil, D., Sri Ganji, R., Bologna, S., et al. (2024). Carboxy-terminal polyglutamylation regulates signaling and phase separation of the Dishevelled protein. EMBO J. *43*, 5635–5666. https://doi.org/10.1038/s44318-024-00254-7.

32. Amacher, J.F., Brooks, L., Hampton, T.H., and Madden, D.R. (2020). Specificity in PDZ-peptide interaction networks: Computational analysis and review. J. Struct. Biol. X *4*, 100022. https://doi.org/10.1016/j.yjsbx.2020.100022.

33. Ernst, A., Appleton, B.A., Ivarsson, Y., Zhang, Y., Gfeller, D., Wiesmann, C., and Sidhu, S.S. (2014). A structural portrait of the PDZ domain family. J. Mol. Biol. *426*, 3509–3519. https://doi.org/10.1016/j.jmb.2014.08.012.

34. Ivarsson, Y., Arnold, R., McLaughlin, M., Nim, S., Joshi, R., Ray, D., Liu, B., Teyra, J., Pawson, T., Moffat, J., et al. (2014). Large-scale interaction profiling of PDZ domains through proteomic peptide-phage display using human and viral phage peptidomes. Proc. Natl. Acad. Sci. USA *111*, 2542–2547. https://doi.org/10.1073/pnas.1312296111.

35. Davey, N.E., Seo, M.H., Yadav, V.K., Jeon, J., Nim, S., Krystkowiak, I., Blikstad, C., Dong, D., Markova, N., Kim, P.M., and Ivarsson, Y. (2017). Discovery of short linear motif-mediated interactions through phage display of intrinsically disordered regions of the human proteome. FEBS J. *284*, 485–498. https://doi.org/10.1111/febs.13995.

36. Ivarsson, Y. (2012). Plasticity of PDZ domains in ligand recognition and signaling. FEBS Lett. *586*, 2638–2647. https://doi.org/10.1016/j.febslet.2012.04.015.

37. Tonikian, R., Zhang, Y., Sazinsky, S.L., Currell, B., Yeh, J.H., Reva, B., Held, H.A., Appleton, B.A., Evangelista, M., Wu, Y., et al. (2008). A specificity map for the PDZ domain family. PLoS Biol. *6*, e239. https://doi.org/10.1371/journal.pbio.0060239.

38. Hillier, B.J., Christopherson, K.S., Prehoda, K.E., Bredt, D.S., and Lim, W. A. (1999). Unexpected modes of PDZ domain scaffolding revealed by structure of nNOS-syntrophin complex. Science *284*, 812–815. https://doi.org/10.1126/SCIENCE.284.5415.812.

39. Penkert, R.R., DiVittorio, H.M., and Prehoda, K.E. (2004). Internal recognition through PDZ domain plasticity in the Par-6-Pals1 complex. Nat. Struct. Mol. Biol. *11*, 1122–1127. https://doi.org/10.1038/nsmb839.

40. Mack, N.A., Porter, A.P., Whalley, H.J., Schwarz, J.P., Jones, R.C., Khaja, A.S.S., Bjartell, A., Anderson, K.I., and Malliri, A. (2012). beta2-syntrophin and Par-3 promote an apicobasal Rac activity gradient at cell-cell junctions by differentially regulating Tiam1 activity. Nat. Cell Biol. *14*, 1169–1180. https://doi.org/10.1038/ncb2608.

41. Gao, C., and Chen, Y.G. (2010). Dishevelled: The hub of Wnt signaling. Cell. Signal. *22*, 717–727. https://doi.org/10.1016/j.cellsig.2009.11.021.

42. Sugi, T., Oyama, T., Muto, T., Nakanishi, S., Morikawa, K., and Jingami, H. (2007). Crystal structures of autoinhibitory PDZ domain of Tamalin: implications for metabotropic glutamate receptor trafficking regulation. EMBO J. *26*, 2192–2205. https://doi.org/10.1038/sj.emboj.7601651.

43. Chang, B.H., Gujral, T.S., Karp, E.S., BuKhalid, R., Grantcharova, V.P., and MacBeath, G. (2011). A systematic family-wide investigation reveals that ~30% of mammalian PDZ domains engage in PDZ-PDZ interactions. Chem. Biol. *18*, 1143–1152. https://doi.org/10.1016/j.chembiol.2011.06.013.

44. Genera, M., Samson, D., Raynal, B., Haouz, A., Baron, B., Simenel, C., Guerois, R., Wolff, N., and Caillet-Saguy, C. (2019). Structural and functional characterization of the PDZ domain of the human phosphatase PTPN3 and its interaction with the human papillomavirus E6 oncoprotein. Sci. Rep. *9*, 7438. https://doi.org/10.1038/s41598-019-43932-x.

45. Fujiwara, Y., Goda, N., Tamashiro, T., Narita, H., Satomura, K., Tenno, T., Nakagawa, A., Oda, M., Suzuki, M., Sakisaka, T., et al. (2015). Crystal structure of afadin PDZ domain-nectin-3 complex shows the structural plasticity of the ligand-binding site. Protein Sci. *24*, 376–385. https://doi.org/10.1002/pro.2628.

46. Ponna, S.K., Ruskamo, S., Myllykoski, M., Keller, C., Boeckers, T.M., and Kursula, P. (2018). Structural basis for PDZ domain interactions in the post-synaptic density scaffolding protein Shank3. J. Neurochem. *145*, 449–463. https://doi.org/10.1111/jnc.14322.

47. Vangone, A., and Bonvin, A.M. (2015). Contacts-based prediction of binding affinity in protein-protein complexes. eLife *4*, e07454. https://doi.org/10.7554/eLife.07454.

48. Appleton, B.A., Zhang, Y., Wu, P., Yin, J.P., Hunziker, W., Skelton, N.J., Sidhu, S.S., and Wiesmann, C. (2006). Comparative structural analysis of the Erbin PDZ domain and the first PDZ domain of ZO-1. Insights into determinants of PDZ domain specificity. J. Biol. Chem. *281*, 22312–22320. https://doi.org/10.1074/jbc.M602901200.

49. Marcotte, D.J., Hus, J.C., Banos, C.C., Wildes, C., Arduini, R., Bergeron, C., Hession, C.A., Baker, D.P., Lin, E., Guckian, K.M., et al. (2018). Lock and chop: A novel method for the generation of a PICK1 PDZ domain and piperidine-based inhibitor co-crystal structure. Protein Sci. *27*, 672–680. https://doi.org/10.1002/pro.3361.

50. Tripsianes, K., Chu, N.K., Friberg, A., Sattler, M., and Becker, C.F.W. (2014). Studying weak and dynamic interactions of posttranslationally modified proteins using expressed protein ligation. ACS Chem. Biol. *9*, 347–352. https://doi.org/10.1021/cb400723j.

51. Papageorgiou, A.C., Pospisilova, M., Cibulka, J., Ashraf, R., Waudby, C. A., Kadeřávek, P., Maroz, V., Kubicek, K., Prokop, Z., Krejci, L., and Tripsianes, K. (2023). Recognition and coacervation of G-quadruplexes by a multifunctional disordered region in RECQ4 helicase. Nat. Commun. *14*, 6751. https://doi.org/10.1038/s41467-023-42503-z.

52. Egea-Jimenez, A.L., Gallardo, R., Garcia-Pino, A., Ivarsson, Y., Wawrzyniak, A.M., Kashyap, R., Loris, R., Schymkowitz, J., Rousseau, F., and Zimmermann, P. (2016). Frizzled 7 and PIP2 binding by syntenin PDZ2 domain supports Frizzled 7 trafficking and signalling. Nat. Commun. *7*, 12101. https://doi.org/10.1038/ncomms12101.

53. Ivarsson, Y., Wawrzyniak, A.M., Kashyap, R., Polanowska, J., Betzi, S., Lembo, F., Vermeiren, E., Chiheb, D., Lenfant, N., Morelli, X., et al. (2013). Prevalence, specificity and determinants of lipid-interacting PDZ domains from an in-cell screen and in vitro binding experiments. PLoS One *8*, e54581. https://doi.org/10.1371/journal.pone.0054581.

54. Liu, X., and Fuentes, E.J. (2019). Emerging Themes in PDZ Domain Signaling: Structure, Function, and Inhibition. Int. Rev. Cell Mol. Biol. *343*, 129–218. https://doi.org/10.1016/bs.ircmb.2018.05.013.

55. Sundell, G.N., Arnold, R., Ali, M., Naksukpaiboon, P., Orts, J., Güntert, P., Chi, C.N., and Ivarsson, Y. (2018). Proteome-wide analysis of phospho-regulated PDZ domain interactions. Mol. Syst. Biol. *14*, e8129. https://doi.org/10.15252/msb.20178129.

Please cite this article in press as: Kumar et al., A class III ligand oscillates between internal and terminal binding modes as it engages with the Dishevelled PDZ domain, Structure (2025), https://doi.org/10.1016/j.str.2025.05.012

56. Gerek, Z.N., and Ozkan, S.B. (2011). Change in allosteric network affects binding affinities of PDZ domains: analysis through perturbation response scanning. PLoS Comput. Biol. *7*, e1002154. https://doi.org/10.1371/journal.pcbi.1002154.

57. Kumawat, A., and Chakrabarty, S. (2017). Hidden electrostatic basis of dynamic allostery in a PDZ domain. Proc. Natl. Acad. Sci. USA *114*, E5825–E5834. https://doi.org/10.1073/pnas.1705311114.

58. Petit, C.M., Zhang, J., Sapienza, P.J., Fuentes, E.J., and Lee, A.L. (2009). Hidden dynamic allostery in a PDZ domain. Proc. Natl. Acad. Sci. USA *106*, 18249–18254. https://doi.org/10.1073/pnas.0904492106.

59. Manjunath, G.P., Ramanujam, P.L., and Galande, S. (2018). Structure function relations in PDZ-domain-containing proteins: Implications for protein networks in cellular signalling. J. Biosci. *43*, 155–171.

60. Tripsianes, K., Madl, T., Machyna, M., Fessas, D., Englbrecht, C., Fischer, U., Neugebauer, K.M., and Sattler, M. (2011). Structural basis for dimethylarginine recognition by the Tudor domains of human SMN and SPF30 proteins. Nat. Struct. Mol. Biol. *18*, 1414–1420. https://doi.org/10.1038/nsmb.2185.

61. Munz, M., Hein, J., and Biggin, P.C. (2012). The role of flexibility and conformational selection in the binding promiscuity of PDZ domains. PLoS Comput. Biol. *8*, e1002749. https://doi.org/10.1371/journal.pcbi.1002749.

62. Haq, S.R., Chi, C.N., Bach, A., Dogan, J., Engström, Å., Hultqvist, G., Karlsson, O.A., Lundström, P., Montemiglio, L.C., Strømgaard, K., et al. (2012). Side-chain interactions form late and cooperatively in the binding reaction between disordered peptides and PDZ domains. J. Am. Chem. Soc. *134*, 599–605. https://doi.org/10.1021/ja209341w.

63. Chi, C.N., Bach, A., Strømgaard, K., Gianni, S., and Jemth, P. (2012). Ligand binding by PDZ domains. Biofactors *38*, 338–348. https://doi.org/10.1002/biof.1031.

64. Ma, W., Chen, M., Kang, H., Steinhart, Z., Angers, S., He, X., and Kirschner, M.W. (2020). Single-molecule dynamics of Dishevelled at the plasma membrane and Wnt pathway activation. Proc. Natl. Acad. Sci. USA *117*, 16690–16701. https://doi.org/10.1073/pnas.1910547117.

65. Rim, E.Y., Clevers, H., and Nusse, R. (2022). The Wnt Pathway: From Signaling Mechanisms to Synthetic Modulators. Annu. Rev. Biochem. *91*, 571–598. https://doi.org/10.1146/annurev-biochem-040320-103615.

66. Bowin, C.F., Kozielewicz, P., Grätz, L., Kowalski-Jahn, M., Schihada, H., and Schulte, G. (2023). WNT stimulation induces dynamic conformational changes in the Frizzled-Dishevelled interaction. Sci. Signal. *16*, eabo4974. https://doi.org/10.1126/scisignal.abo4974.

67. Gammons, M.V., Renko, M., Johnson, C.M., Rutherford, T.J., and Bienz, M. (2016). Wnt Signalosome Assembly by DEP Domain Swapping of Dishevelled. Mol. Cell *64*, 92–104. https://doi.org/10.1016/j.molcel.2016.08.026.

68. Bernatik, O., Sedova, K., Schille, C., Ganji, R.S., Cervenka, I., Trantirek, L., Schambony, A., Zdrahal, Z., and Bryja, V. (2014). Functional analysis of dishevelled-3 phosphorylation identifies distinct mechanisms driven by casein kinase 1ϵ and frizzled5. J. Biol. Chem. *289*, 23520–23533. https://doi.org/10.1074/jbc.M114.590638.

69. Bikkavilli, R.K., Avasarala, S., Vanscoyk, M., Sechler, M., Kelley, N., Malbon, C.C., and Winn, R.A. (2012). Dishevelled3 is a novel arginine methyl transferase substrate. Sci. Rep. *2*, 805. https://doi.org/10.1038/srep00805.

70. Qi, J., Lee, H.J., Saquet, A., Cheng, X.N., Shao, M., Zheng, J.J., and Shi, D.L. (2017). Autoinhibition of Dishevelled protein regulated by its extreme C terminus plays a distinct role in Wnt/beta-catenin and Wnt/planar cell polarity (PCP) signaling pathways. J. Biol. Chem. *292*, 5898–5908. https://doi.org/10.1074/jbc.M116.772509.

71. Hanakova, K., Bernatik, O., Kravec, M., Micka, M., Kumar, J., Harnos, J., Ovesna, P., Paclikova, P., Radsetoulal, M., Potesil, D., et al. (2019). Comparative phosphorylation map of Dishevelled 3 links phospho-signatures to biological outputs. Cell Commun. Signal. *17*, 170. https://doi.org/10.1186/s12964-019-0470-z.

72. Bui, T.D., Beier, D.R., Jonssen, M., Smith, K., Dorrington, S.M., Kaklamanis, L., Kearney, L., Regan, R., Sussman, D.J., and Harris, A.L. (1997). cDNA cloning of a human dishevelled DVL-3 gene, mapping to 3q27, and expression in human breast and colon carcinomas. Biochem. Biophys. Res. Commun. *239*, 510–516. https://doi.org/10.1006/bbrc.1997.7500.

73. Mizutani, K., Miyamoto, S., Nagahata, T., Konishi, N., Emi, M., and Onda, M. (2005). Upregulation and overexpression of DVL1, the human counterpart of the Drosophila dishevelled gene. Tumori *91*, 546–551. https://doi.org/10.1177/030089160509100616.

74. Uematsu, K., Kanazawa, S., You, L., He, B., Xu, Z., Li, K., Peterlin, B.M., McCormick, F., and Jablons, D.M. (2003). Wnt pathway activation in mesothelioma: evidence of Dishevelled overexpression and transcriptional activity of beta-catenin. Cancer Res. *63*, 4547–4551.

75. Zhang, C., Li, C., Chen, X., Zhou, Y., Yin, B., Ni, R., Zhang, Y., and Liu, J. (2017). Overexpression of dishevelled 2 is involved in tumor metastasis and is associated with poor prognosis in hepatocellular carcinoma. Clin. Transl. Oncol. *19*, 1507–1517. https://doi.org/10.1007/s12094-017-1697-z.

76. Castro-Piedras, I., Sharma, M., Brelsfoard, J., Vartak, D., Martinez, E.G., Rivera, C., Molehin, D., Bright, R.K., Fokar, M., Guindon, J., and Pruitt, K. (2021). Nuclear Dishevelled targets gene regulatory regions and promotes tumor growth. EMBO Rep. *22*, e50600. https://doi.org/10.15252/embr.202050600.

77. Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191. https://doi.org/10.1093/bioinformatics/btp033.

78. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340.

79. Krug, M., Weiss, M.S., Heinemann, U., and Mueller, U. (2012). a graphical user interface for the convenient processing of diffraction data using. J. Appl. Crystallogr. *45*, 568–572. https://doi.org/10.1107/S0021889812011715.

80. Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A., et al. (2011). Overview of the CCP4 suite and current developments. Acta Crystallogr. D Biol. Crystallogr. *67*, 235–242. https://doi.org/10.1107/S0907444910045749.

81. McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. (2007). Phaser crystallographic software. J. Appl. Crystallogr. *40*, 658–674. https://doi.org/10.1107/S0021889807021206.

82. Murshudov, G.N., Vagin, A.A., and Dodson, E.J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr. D Biol. Crystallogr. *53*, 240–255. https://doi.org/10.1107/S0907444996012255.

83. Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. *60*, 2126–2132. https://doi.org/10.1107/S0907444904019158.

84. Gore, S., Sanz García, E., Hendrickx, P.M.S., Gutmanas, A., Westbrook, J.D., Yang, H., Feng, Z., Baskaran, K., Berrisford, J.M., Hudson, B.P., et al. (2017). Validation of Structures in the Protein Data Bank. Structure *25*, 1916–1927. https://doi.org/10.1016/j.str.2017.10.009.

85. Manalastas-Cantos, K., Konarev, P.V., Hajizadeh, N.R., Kikhney, A.G., Petoukhov, M.V., Molodenskiy, D.S., Panjkovich, A., Mertens, H.D.T., Gruzinov, A., Borges, C., et al. (2021). ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. J. Appl. Crystallogr. *54*, 343–355. https://doi.org/10.1107/S1600576720013412.

86. Lee, W., Tonelli, M., and Markley, J.L. (2015). NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. Bioinformatics *31*, 1325–1327. https://doi.org/10.1093/bioinformatics/btu830.

87. Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., and Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based

Please cite this article in press as: Kumar et al., A class III ligand oscillates between internal and terminal binding modes as it engages with the Dishevelled PDZ domain, Structure (2025), https://doi.org/10.1016/j.str.2025.05.012

on UNIX pipes. J. Biomol. NMR *6*, 277–293. https://doi.org/10.1007/BF00197809.

88. Tripsianes, K., Schütz, U., Emmanouilidis, L., Gemmecker, G., and Sattler, M. (2019). Selective isotope labeling for NMR structure determination of proteins in complex with unlabeled ligands. J. Biomol. NMR *73*, 183–189. https://doi.org/10.1007/s10858-019-00241-9.

89. Lacabanne, D., Meier, B.H., and Böckmann, A. (2018). Selective labeling and unlabeling strategies in protein solid-state NMR spectroscopy. J. Biomol. NMR *71*, 141–150. https://doi.org/10.1007/s10858-017-0156-z.

90. Sattler, M., Schleucher, J., and Griesinger, C. (1999). Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. Prog Nucl Mag Res Sp *34*, 93–158. https://doi.org/10.1016/S0079-6565(98)00025-9.

91. Evangelidis, T., Nerli, S., Nováček, J., Brereton, A.E., Karplus, P.A., Dotas, R.R., Venditti, V., Sgourakis, N.G., and Tripsianes, K. (2018). Automated NMR resonance assignments and structure determination using a minimal set of 4D spectra. Nat. Commun. *9*, 384. https://doi.org/10.1038/s41467-017-02592-z.