Ankyrin repeats in context with human population variation

Utgés, Javier S.; Tsenkov, Maxim I.; Dietrich, Noah J. M.; MacGowan, Stuart A.; Barton, Geoffrey J.

DOI:
10.1101/2021.05.28.445974

Publication date:
2021

Document Version
Other version

Ankyrin repeats in context with human population variation

Javier S. Utgés[1,2], Maxim I. Tsenkov[1], Noah J. M. Dietrich[1], Stuart A. MacGowan[1] and Geoffrey J. Barton[1*]

[1]Division of Computational Biology, School of Life Sciences, University of Dundee, Scotland, UK

[2]Universitat Pompeu Fabra (UPF), Barcelona, Spain

[*]Correspondence to: gjbarton@dundee.ac.uk

Abstract

Ankyrin protein repeats bind to a wide range of substrates and are one of the most common protein motifs in nature. Here, we collate a high-quality alignment of 7,407 ankyrin repeats and examine for the first time, the distribution of human population variants from large-scale sequencing of healthy individuals across this family. Population variants are not randomly distributed across the genome but are constrained by gene essentiality and function. Accordingly, we interpret the population variants in context with evolutionary constraint and structural features including secondary structure, accessibility and protein-protein interactions across 383 three-dimensional structures of ankyrin repeats. We find five positions that are highly conserved across homologs and also depleted in missense variants within the human population. These positions are significantly enriched in intra-domain contacts and so likely to be key for repeat packing. In contrast, a group of evolutionarily divergent positions are found to be depleted in missense variants in human but significantly enriched in protein-protein interactions. Our analysis also suggests the domain has three, not two surfaces, each with different patterns of enrichment in protein-substrate interactions and missense variants. Our findings will be of interest to those studying or engineering ankyrin-repeat containing proteins as well as those interpreting the significance of disease variants.

Author Summary

Comparison of variation at each position of the amino acid sequence for a protein across different species is a powerful way to identify parts of the protein that are important for its structure and function. Large-scale DNA sequencing of healthy people has recently made it possible to study normal genetic variation within just one species. Our work combines information on genetic differences between over 100,000 people with in-depth analysis of all available three-dimensional structures for Ankyrin repeats which are a widespread family of binding proteins formed by units with similar amino acid sequence that are found in tandem. Our combined analysis identifies sites critical for ankyrin stability as well as the positions most important for substrate interactions and hence function. Although focused only on the Ankyrins, the principles developed in our work are general and can be applied to any protein family.

Introduction

The ankyrin repeat motif (ANK) is one of the most commonly observed protein motifs in nature, with proteins containing this motif found in practically all phyla (1). Ankyrin repeats (AR) are specialised in protein binding and take part in many processes including transcription initiation, cell cycle regulation and cell signalling (2). ANK is 33 residues long (Fig 1A) and has a helix-turn-helix conformation, with short loops at the N and C termini (Fig 1B). The last and first two residues of adjacent repeats form a β-turn. These β-turns project outward at an angle of ≈ 90° to the antiparallel α-helices, yielding the characteristic L-shaped cross section of ankyrin repeats. Ankyrin repeats are usually found in tandem with two or more forming an ankyrin repeat domain (ARD). The stacking of repeats is mediated by the conserved hydrophobic faces of the helices as well as the complementarity of repeat surfaces that assemble to form an extended helical bundle (Fig 1C) (3). Less conserved positions in the motif, i.e., positions that present residues with many different physicochemical properties, are located on the surface, and are likely to interact with ligands. More conserved positions, on the other hand, tend to be buried in the structure and are responsible for the correct packing of the domain. They do this by forming both intra- and inter-repeat contacts, such as hydrophobic and hydrogen bond interactions (4).

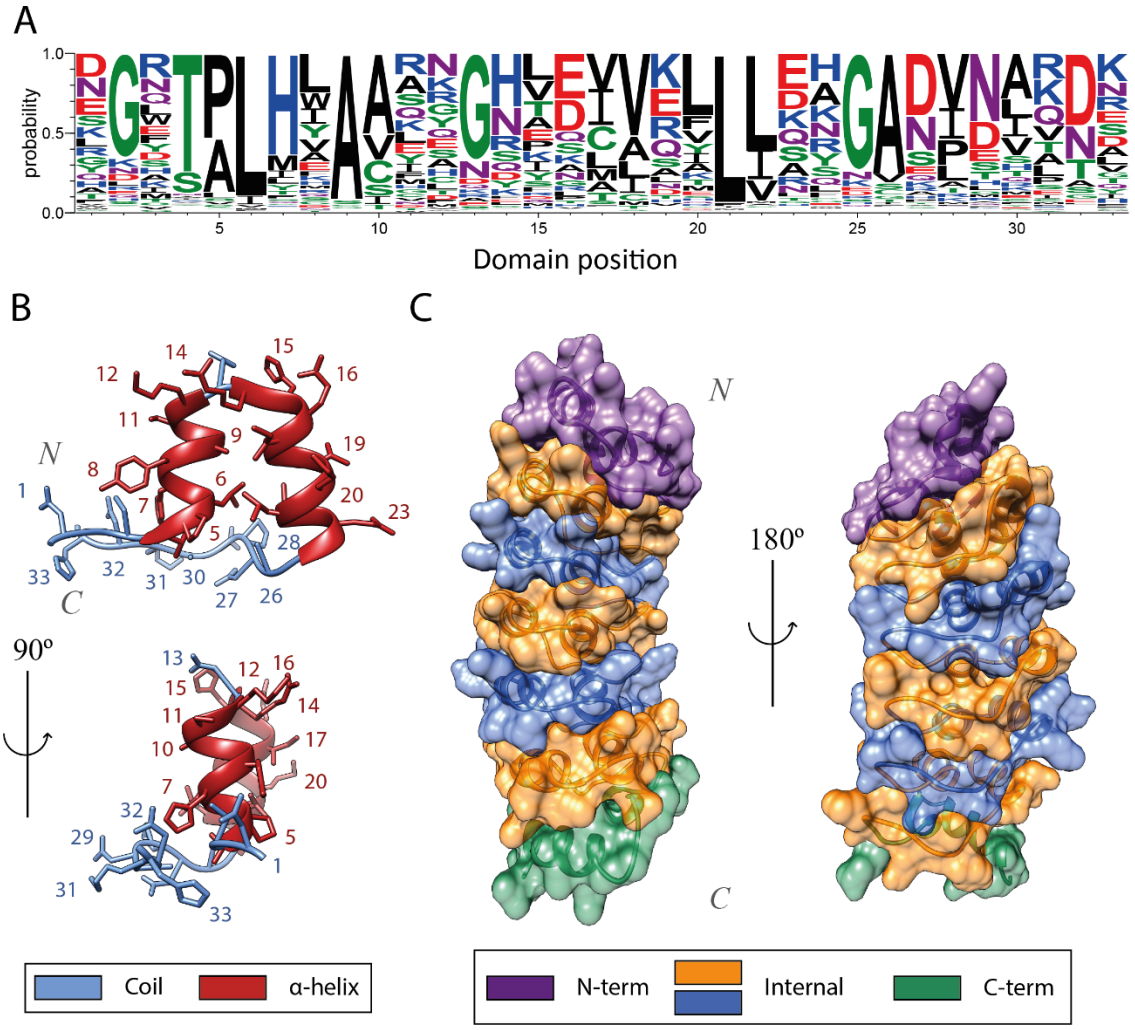

Figure 1. (A) Sequence logo of the ANK obtained with WebLogo (5) derived from the MSA generated in this work. The Y axis indicates the probability of observing an amino acid at any position within the motif; (B) Tertiary structure of an ankyrin repeat, coloured by secondary structure class: helices in red and coil in blue; (C) Representation of the complementary surfaces of individual ARs that form the human gankyrin ARD surface. N- and C-capping AR surfaces are coloured in purple and green respectively, whereas modular ones are coloured in blue and orange. (PDB ID: 1UOH) (6). Structure visualization with UCSF Chimera (7).

Proteins containing ankyrin repeats are known to bind many different protein and small molecule substrates. The concave face of an ARD, comprising the β-turn/loop region and the first α-helix is often associated with substrate binding. (8). Recent evidence suggests that ARDs might not only be able to bind small ligands or proteins, but also a range of sugars and lipids, thus extending their versatility and flexibility in substrate binding (9). This, coupled with the success that designed ankyrin repeat proteins (DARPins) are having in the clinical field (10) make ankyrins an extremely interesting target to study.

Within the sequence variability found in protein repeats, ankyrins are relatively conserved and multiple amino acid patterns can be observed within the ANK (11) (Fig 1A). The TPLH motif, positions 4-7, is highly conserved across all ankyrin repeats. It is found at the beginning of the first α-helix. Thr4 establishes three hydrogen bonds with His7 (Fig 2E), Pro5 starts the helix with a tight turn and Leu6 forms multiple hydrophobic interactions both within and between

repeats (Fig 2B). The loops are more diverse in sequence, yet certain patterns are apparent as well. The subsequence GADVN, 25-29, can be observed in the loop connecting the second α-helix with the β-turn. Gly25 breaks the second helix. Ala26 and Val28 form intra- and inter-repeat interactions (Fig 2B), whereas Asp27 and Asn29 form hydrogen bonds with adjacent repeats (Fig 2D). Gly13 is found in the loop between the antiparallel helices. Asp32 and Gly2 are highly conserved at the β-turn that connects repeats (Fig 2C). A total of six hydrogen bonds take place in this turn. This explains why Asp32 is conserved, and Gly2 would be similar to Gly13 and Gly25, and is conserved due to its flexibility and special structural features (12). In the second α-helix, the [I/V]VXLLL hydrophobic motif is observed. The residues on this motif, except X19, which is usually hydrophilic, form intra- and inter-repeat hydrophobic networks that are thought to help keep together the ARD structure (13) (Fig 2B).

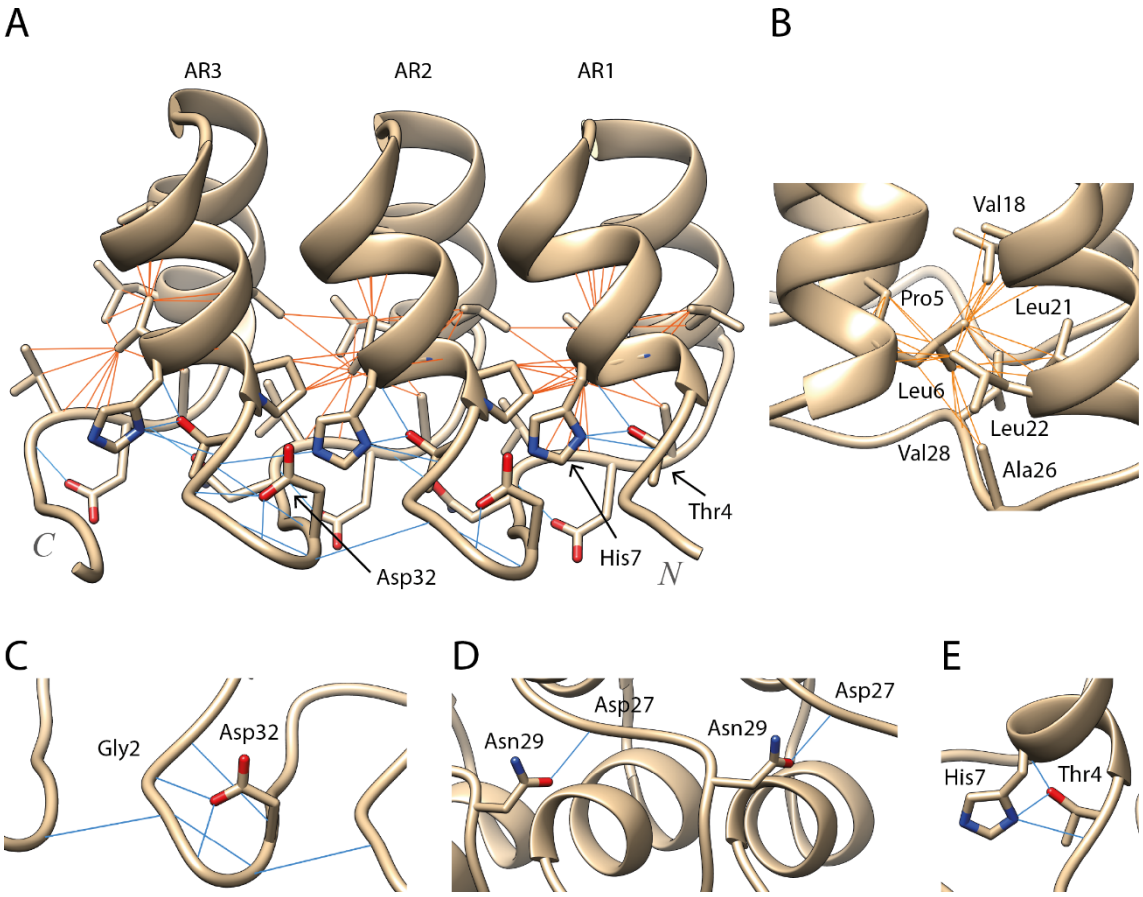

Figure 2. (A) Trio of ARs from a designed ankyrin repeat protein (14) (PDB: 5MA3). These three ARs display the main known interactions responsible for the correct packing of the ARD. Hydrogen bond interactions are depicted by blue lines whereas hydrophobic ones are depicted by orange ones; (B) Hydrophobic network formed by Leu6 in the hydrophobic core of the domain; (C) Hydrogen bonding network at the β-turn between positions Asp32-Gly2; (D) Inter-repeat hydrogen bonds between conserved Asn29 and Asp27; (E) Thr4 forms three hydrogen bonds with His7. Structure visualization with UCSF Chimera (7).

In protein sequences, functional or structurally important residues are constrained in evolution, resulting in amino acid conservation between homologues. In a similar way, the genomic distribution of genetic variation within a species is affected by factors such as gene essentiality (15), and protein domain architecture (16). Synonymous variants do not change the protein sequence, and as a result they appear randomly distributed in structure. Missense variants, on the other hand, change the protein sequence and are consequently constrained in space.

Pathogenic missense variants cluster in three-dimensional structure around functionally important regions, such as catalytic sites, whereas neutral or non-pathogenic missense variants tend to aggregate on regions which are tolerant to amino acid substitutions (17). Recently, in a study of all Pfam domain families, MacGowan, Madeira (18) found that positions conserved across homologues but also depleted in missense variants within the human population were of particular functional and/or structural relevance since they are heavily enriched in disease-associated variants. They also found a subset of evolutionary unconserved positions that were missense depleted. These positions were enriched in ligand, DNA and protein interactions as well as in pathogenic variants, suggesting their functional importance within the protein domain.

In this paper we perform a novel analysis that combines human population genetic variation from gnomAD (19) across ankyrin repeats in context with evolutionary variation and all available ankyrin protein structures. This is the first in-depth application to a repeat family of the concepts developed in our earlier work across all Pfam domains (18). Application to a repeat family boosts the statistical power of the method and highlights the positions in the ANK most likely to be important for structural stability as well as those relevant to substrate specificity. We anticipate that this work will be of value to those interested in understanding the function of ANK containing proteins as well as those aiming to engineer novel AR specificity.

Methods

Sequence extraction and database integration

The UniProt (20), SMART (SM00248) (21), ProSite (PS50088) (22), PRINTS (PR01415) (23) and PFAM (PF13606, PF00023) (24) were scanned for ankyrin repeat motif (ANK) definitions in all species. The databases use slightly different algorithms resulting in variation in the number as well as the length and coordinates of annotations between them (Fig 3A). Accordingly, we retrieved all ankyrin repeat sequences found in Swiss-Prot reviewed proteins from the following databases: UniProt (7,230 ankyrin repeats), SMART (6,396), ProSite (4,119), PRINTS (796) and PFAM (288) (Fig 3B) resulting in a total of 18,825 ANK annotations. After redundancy filtering, we established a high-quality set of 7,407 ankyrin repeat sequences: 4,109 (ProSite), 2,313 (SMART), 972 (UniProt) and 10 (PFAM) (Fig 3C) for analysis.

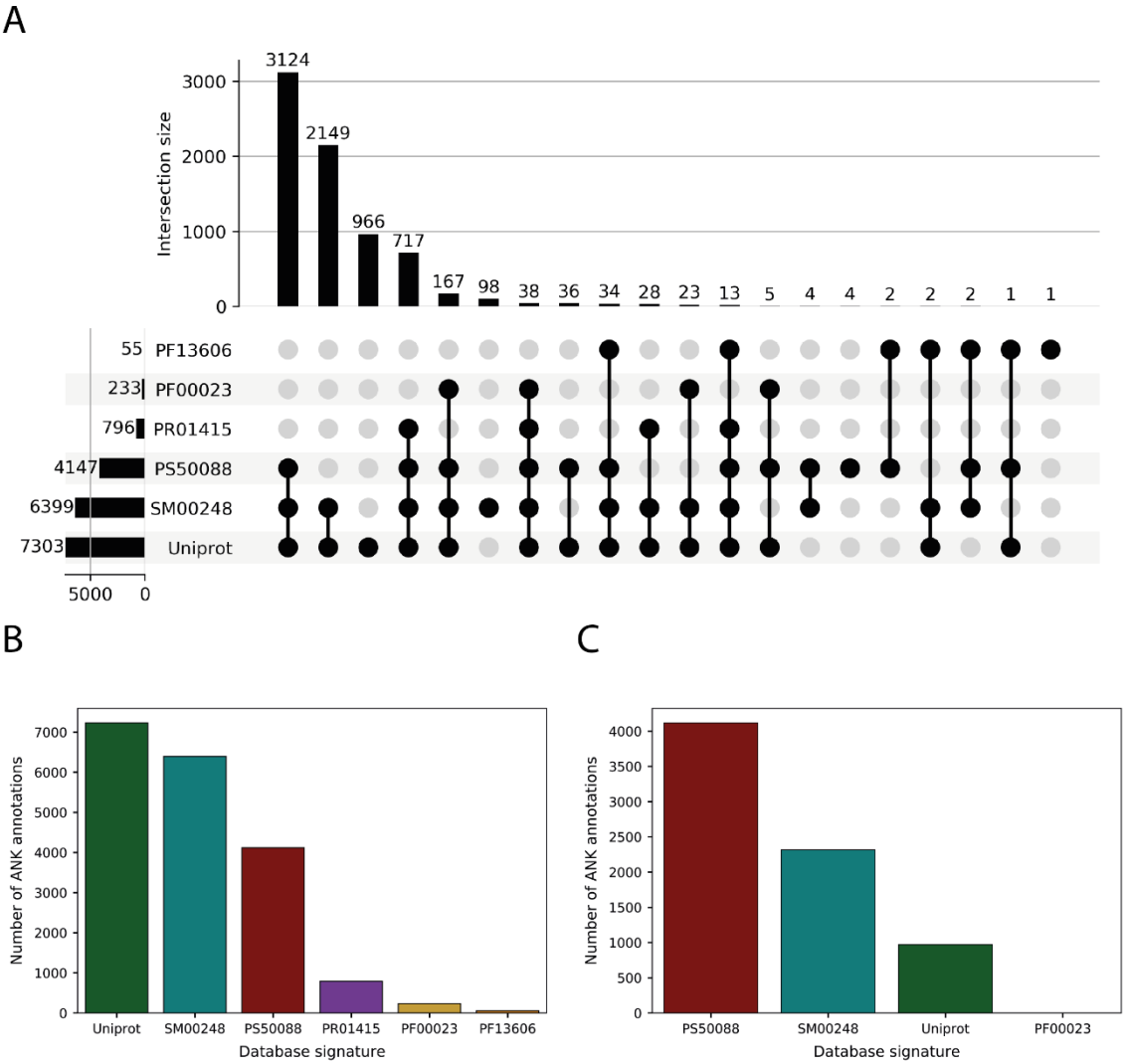

Figure 3. (A) Upset plot showing the distribution of ANK annotations and the overlap between different database signatures. Most of the annotations are shared between UniProt, SM00248 and PS50088. UniProt presents ≈ 1000 unique annotations which are not present in any other database; (B) This bar plot indicates the number of ANK annotations per database signature: 7,230, 6,396, 4,119, 796, 233 and 55 from left to right; (C) This bar plot shows the composition of the dataset resulting from the database merging, with ProSite accounting for ≈ 55% of the annotations, SMART for ≈ 30% and UniProt for the last ≈ 15%.

Multiple sequence alignment

Several approaches were tried to align the 7,407 ankyrin repeat (AR) sequences, both sequence and structure-based. These included ClustalΩ (25), HMMER (26), T-Coffee (27), AMPS (28), Muscle (29) and STAMP (30). When applied to all 7,407 sequences, these aligners introduced many gaps and a high proportion of misaligned residues which were inconsistent with known key residues in the ankyrin repeat. Accordingly, the final multiple sequence alignment (MSA) was obtained by carrying out a series of sequences-to-profile multiple sequence alignments with ClustalΩ (Fig 4) as follows.

First, the sequences were divided into different groups according to their length and database of origin. Then, sequences that had the most common length, 33 residues, coming from the highest confidence database, ProSite, were aligned using ClustalΩ version 1.2.2 with defaults. Sequences introducing gaps in the 33 high-occupancy columns were removed and re-aligned with

a ClustalΩ sequences-to-profile alignment. Sequences inserting gaps yet again were removed from the alignment.

For the rest of sequence groups, defined by sequence length and database of origin were aligned to this growing alignment by consecutive sequence-to-profile alignments. As with the first alignment, gap-introducing sequences were re-aligned and removed if necessary, for each group.

At the end of this process, ≈ 98% of the sequences were aligned in the resulting MSA. The remaining 2% was formed by those gap-introducing sequences removed during the re-alignment phase of the process. This 2% of sequences were re-aligned to the main alignment by a profile-to-profile alignment shown as an overview in Figure 4.

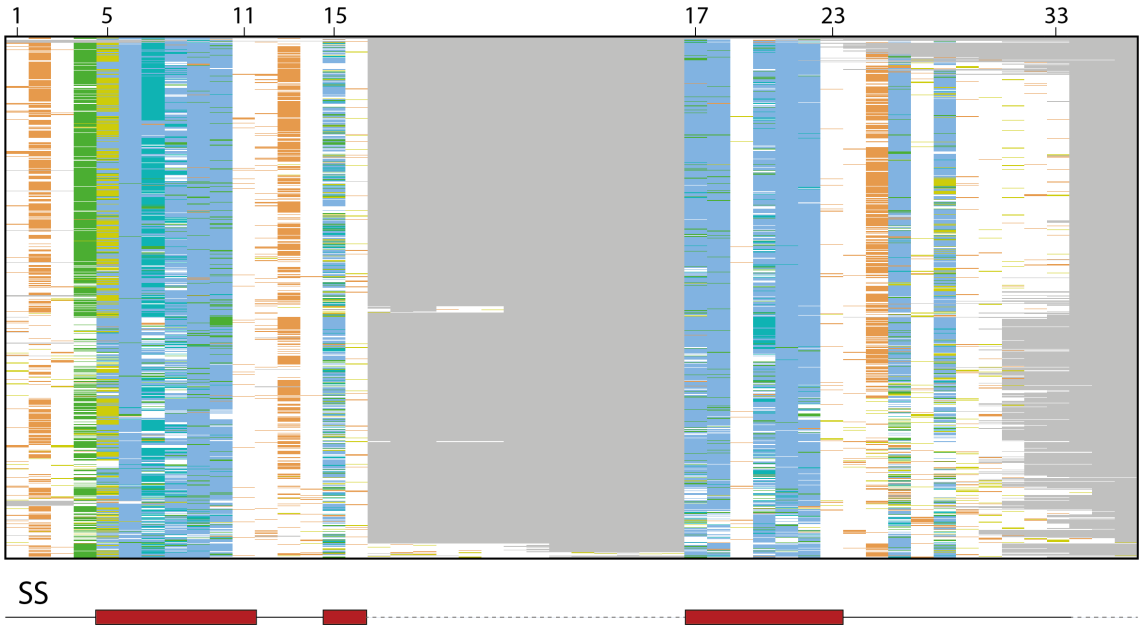

Figure 4. Overview of the resulting MSA, including the 7,404 ankyrin repeat sequences. Only columns presenting an occupancy > 0.5% are shown. Sequences are sorted by a tree generated in Jalview using the average distance method and the BLOSUM62 matrix. Columns between 16-17 and after 33 represent insertions in some ankyrin repeats. Red boxes below the overview indicate the location of the secondary structure elements (SS), α-helices in this case, within the alignment. Grey dashed lines represent gaps and are mostly found at low-occupancy columns. Columns are coloured according to the ClustalX colour scheme (31). Hydrophobic residues are coloured in blue, glycines in orange, prolines in yellow, polar residues in green and unconserved columns are coloured in white. Obtained with Jalview (32).

VarAlign and ProIntVar

A total of 35,691 variants found in the genome aggregation database (gnomAD) (19) coming from 1,435 human sequences were mapped to the MSA through VarAlign (18) (Fig 5).

419 sequences in the alignment were mapped to 209 different structures solved by X-ray crystallography in the PDB (33-35) via SIFTS (36) through ProIntVar (37). These sequences correspond to 419 unique ankyrin repeats, found in 80 proteins. The real-space R value (RSR) and RSR-Z scores, as well as the real-space correlation coefficient (RSCC) quality metrics, as calculated by (38), were retrieved by ProteoFAV (18) from the validation reports in PDBe. Only residues with RSCC > 0.85 and RSRZ < 2 were considered for analysis. After this filtering step, our structural dataset comprised 383/419 unique ARs coming from 176/209 PDBs, representing 73/80 proteins. This dataset included 11,186 of the 13,059 residues with structural coverage

before quality filtering. The average RSRZ per residue after filtering was −0.11 and the mean RSCC had a value of +0.95.

DSSP (39) was run on all structures via ProIntVar and information from 381 ankyrin repeat sequences was used to determine the consensus secondary structure as well as the relative solvent accessibility (RSA) classification for all positions in the ANK, as described in MacGowan, Madeira (18).

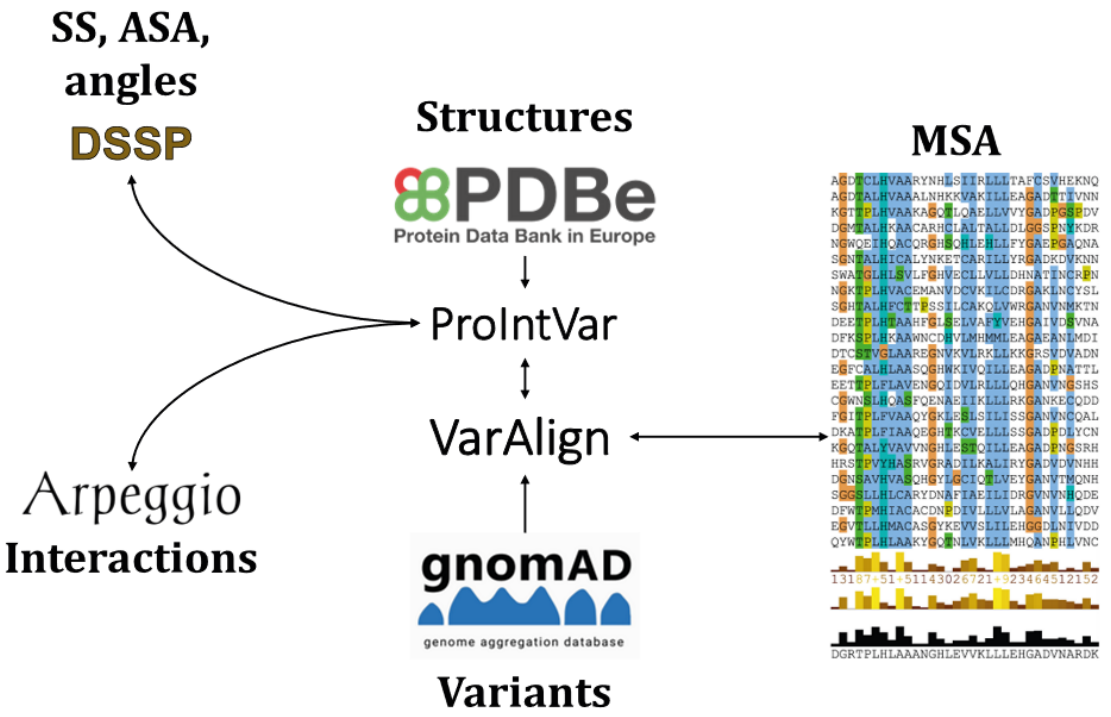

Figure 5. Diagram showing the main components of the pipeline. VarAlign retrieves variants found in human sequences in the MSA from gnomAD. ProIntVar retrieves structures from the PDBe and runs DSSP and Arpeggio to get secondary structure, accessible surface area and inter-atomic contacts information. Everything is mapped back to the residues and MSA columns (37).

Sequence divergence score

The Shenkin divergence score was used to characterise residue conservation at an alignment position (40). This is a divergence score, based on Shannon's entropy (Equations 1 and 2).

$$V_{Shenkin} = 2^S \times 6 \quad (1)$$

$$S = -\sum_{i}^{K} p_i \log_2 p_i \quad (2)$$

Where S is Shannon's entropy and i is every one of the $K = 20$ different amino acid types. The range of this diversity score is determined by Shannon's entropy. In a completely conserved alignment column, one amino acid residue will be found with a frequency of 1.0, whereas the rest will not be present, resulting in an entropy of 0.0, and a minimum $V_{Shenkin} = 2^0 \times 6 = 6$. At the other extreme, an alignment column with all 20 amino acids at a frequency of 1/20 = 0.05 would give an entropy of $S \approx 4.32$, resulting in a maximum $V_{Shenkin} = 2^{4.32} \times 6 \approx 120$. Thus, low Shenkin scores indicate higher conservation at a position and *vice versa*. To simplify the interpretation of the score, we normalised the Shenkin score to 0-100 (Equation 3).

$$N_{Shenkin} = (V_{Shenkin} - V_{Shenkin_{min}})/(V_{Shenkin_{max}} - V_{Shenkin_{min}}) \quad (3)$$

Where $V_{Shenkin_{min}}$ is the score of the most conserved column within the alignment, Position 9 with a Shenkin score of 15.43 and $V_{Shenkin_{max}}$ is the score of the most diverse position, Position 3 with a score of 103.96.

Enrichment in variants

The human genetic variants from gnomAD (19) were mapped to the MSA and missense variant enrichment scores (MES) were calculated for the 33 positions of the ANK. MES is expressed as the natural logarithm of an odds ratio (OR) and it represents the enrichment of variants in an alignment column relative to the average for the other columns. Columns were classified as depleted, enriched or neutral according to this MES (18). 95% confidence intervals and p-values were calculated to assess the significance of these ratios (41).

Enrichment in protein-substrate interactions

For the structural analysis, the meaningful biological units were retrieved from PDBe. These are the preferred assemblies for each structure, instead of the asymmetric units, which might not reflect the packing of the protein observed in nature. All inter-atomic contacts were calculated by Arpeggio (42). Atoms were considered to interact if they were within 5Å of each other.

We considered all interactions between an ankyrin repeat and any protein substrate present in the preferred assembly as protein-protein interactions (PPIs). A log enrichment score was calculated for PPIs per position in the motif in a similar manner to MES above. It is referred as protein-protein interaction enrichment score (PPIES). The number of protein-protein interactions per alignment column was normalised by the structural coverage of that column in structures presenting an interaction between an ARD and a bound peptide substrate. We considered that there was evidence of contact between an AR position and a bound peptide substrate if there was at least one inter-atomic contact involving the repeat position and the substrate in at least one of the structures representing the complex.

Enrichment in intra-repeat contacts

A contact map, shown as a 33 × 33 matrix, for the 33 positions in the ANK, was calculated to show how often two positions interact within an AR. Each cell shows the proportion of repeats, where evidence of contact between a given pair of residues has been observed. The absolute frequency is normalised by the coverage of a given pair of residues within a repeat. This intra-repeat contact map is symmetric. Thus, a given cell $c_{i,j} = c_{j,i}$. Contacts between adjacent residues are not shown, resulting in a null diagonal.

Enrichment in intra-repeat contacts per position was calculated. Since the intra-repeat contact matrix is symmetric, the total number of contacts per residue, C_i, was calculated using Equation 4, where $c_{i,j}$ is the absolute frequency of contacts between any two amino acid residues present at positions i and j within the $K = 33$ positions in the ANK. The same approach was used to calculate the total structural coverage per ANK position, O_i, (Equation 5) where $o_{i,j}$ is the absolute frequency of both positions i and j being present in the same repeat.

$$C_i = \sum_{j=1}^{K} c_{i,j} \quad (4)$$

$$O_i = \sum_{j}^{K} o_{i,j} \quad (5)$$

The total number of contacts and coverage of a position were calculated as the sum of all their contacts and coverages, respectively (Equations 6 and 7).

$$O_t = \sum_{i}^{K} O_i \quad (6)$$

$$C_t = \sum_{i}^{K} C_i \quad (7)$$

Enrichment in these contacts was calculated per position in the same fashion as for variants and PPIs. The same analysis was carried out on inter-repeat contacts but was of limited value and so not included in this paper.

Results and Discussion

In this work, 7,407 ankyrin repeat sequences, including both human and other species, were used to build a multiple sequence alignment and conservation profile of the motif. Human genetic variation data coming from 1,435 human ankyrin repeats were used to study the distribution of variation within the motif. Moreover, 176 three-dimensional structures, representing a total of 383 different ankyrin repeats were used to structurally characterise in detail this motif by secondary structure, residue solvent accessibility, intra-domain contacts and protein-protein interactions. For the first time, human population variation data was used to explain the evolutionary constraint acting upon this family of protein repeats, integrating at the same time these data with structure and sequence divergence.

262

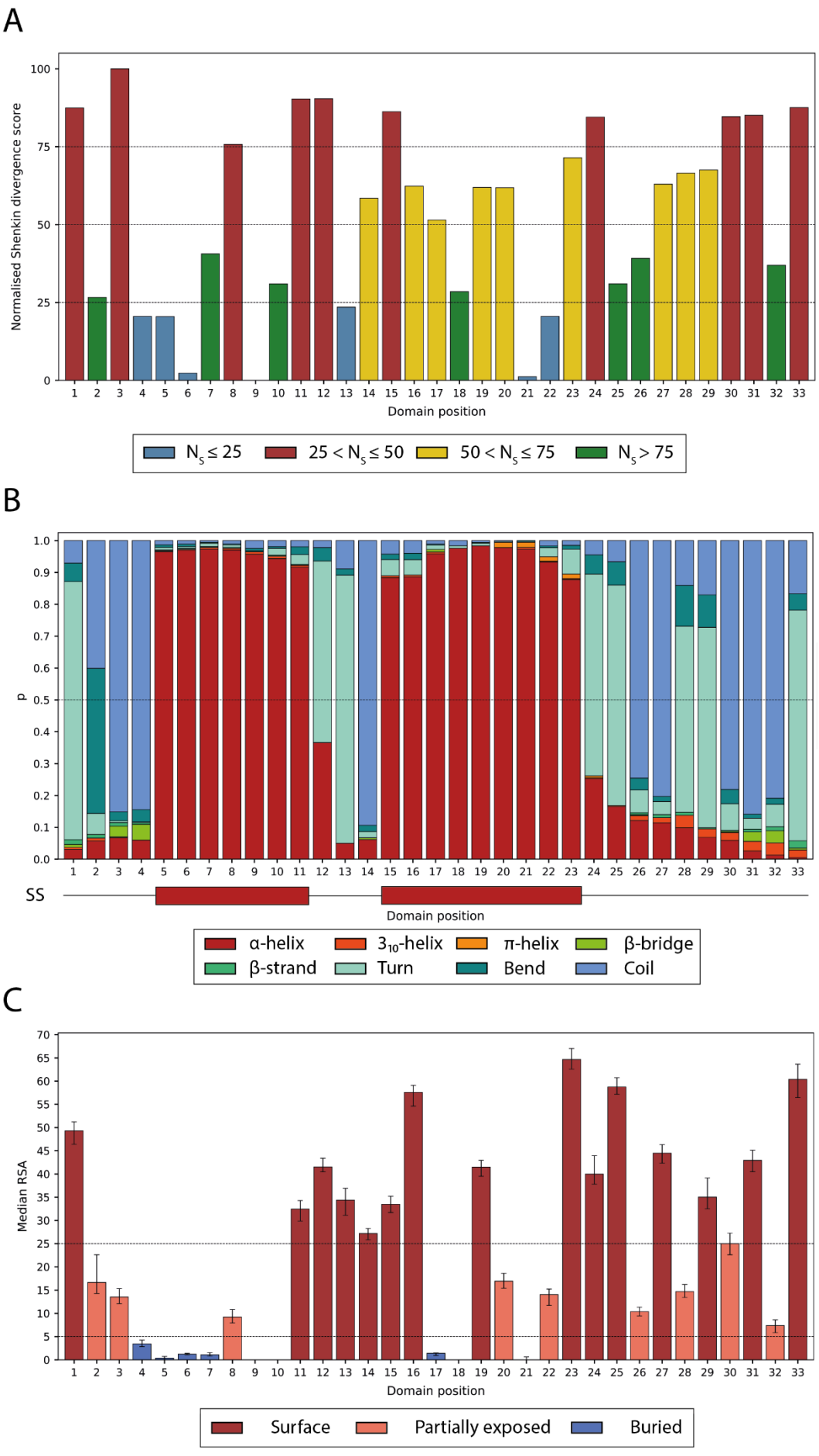

Figure 6. (A) Normalised Shenkin divergence score per domain position (Eq. 3) calculated from the MSA containing 7,404 sequences. Positions are coloured according to their normalised Shenkin score as the legend indicates; (B) Secondary structure assignment per position. Within each position, each coloured bar represents the frequency of the eight states defined by DSSP: α-helix, 3_{10}-helix, π-helix, β-bridge, β-strand, turn, bend and coil, observed for the residues with structural coverage at that column in the MSA. Most helices range from 5-11 and 15-23 and finish in 5-turns, usually at positions 12-13 and 23-24. Two β-turns are observed at positions 28-29 and 33-1; (C) Median residue relative surface accessibility per position, calculated from DSSP′s accessible surface area (39) as described in Tien, Meyer (43). Error bars indicate 95% CI of the median. Positions were classified according to the specified thresholds: surface (RSA ≥ 25%), partially exposed (5% < RSA < 25%) or buried (RSA ≤ 5%) (44).

Conservation profile

The conservation profile derived from our MSA agrees with previous work (13). Figure 6A shows the normalised Shenkin divergence score per position in the motif. As described in Methods, this score goes from 0-100. Among the most conserved positions ($N_{Shenkin} < 25$) we find Thr4, Pro5, Leu6, belonging in the TPLH motif, 4-7 as well as Ala9, Gly13, Leu21 and Leu22. Some of the most evolutionary diverse positions, on the other hand, include positions 1, 3, 11, 12 and 33 among others, all presenting $N_{Shenkin} > 75$. Most of the highly diverse positions are found on the concave surface and contribute to the variable interface where most of the substrate binding takes place.

Secondary structure

Figure 6B shows the secondary structure assignment of the 33 positions in the ANK. Most of the repeats present a seven-residue long first helix ranging from the fifth to the 11th position and a second helix that in most cases is nine-residues long and extends from the 15th to the 23rd position. Our results also show four turns along the ANK. Two of these turns, found at positions 12-13 and 24-25, are 5-turns and simply indicate the end of the α-helices, whereas the other two, positions 28-29 and 33-1, are β-turns. These two β-turns were classified as type I β-turns according to the φ and ψ dihedral angles distribution of consensus columns 33, 1 and 28, 29 (45). Positions 27 and 32 in the alignment present either Asn or Asp with a high frequency of 44% and 58% respectively. Consequently, we classified the turns they initiate as Asx motifs (Fig 7A-C). The turn at positions 27-30 was classified as an Asx-β-turn and the one at 32-2 as a type 1 β-bulge loop with an Asx motif (46). Repeats that do not have Asx at positions 27 and 32, form a simple β-turn, instead of an Asx motif since they lack the extra hydrogen bonds that this secondary structure motif requires. The conservation of these Asx residues on both turns, suggests a structural relevance and role of these Asx motifs on the correct packing of the ARD.

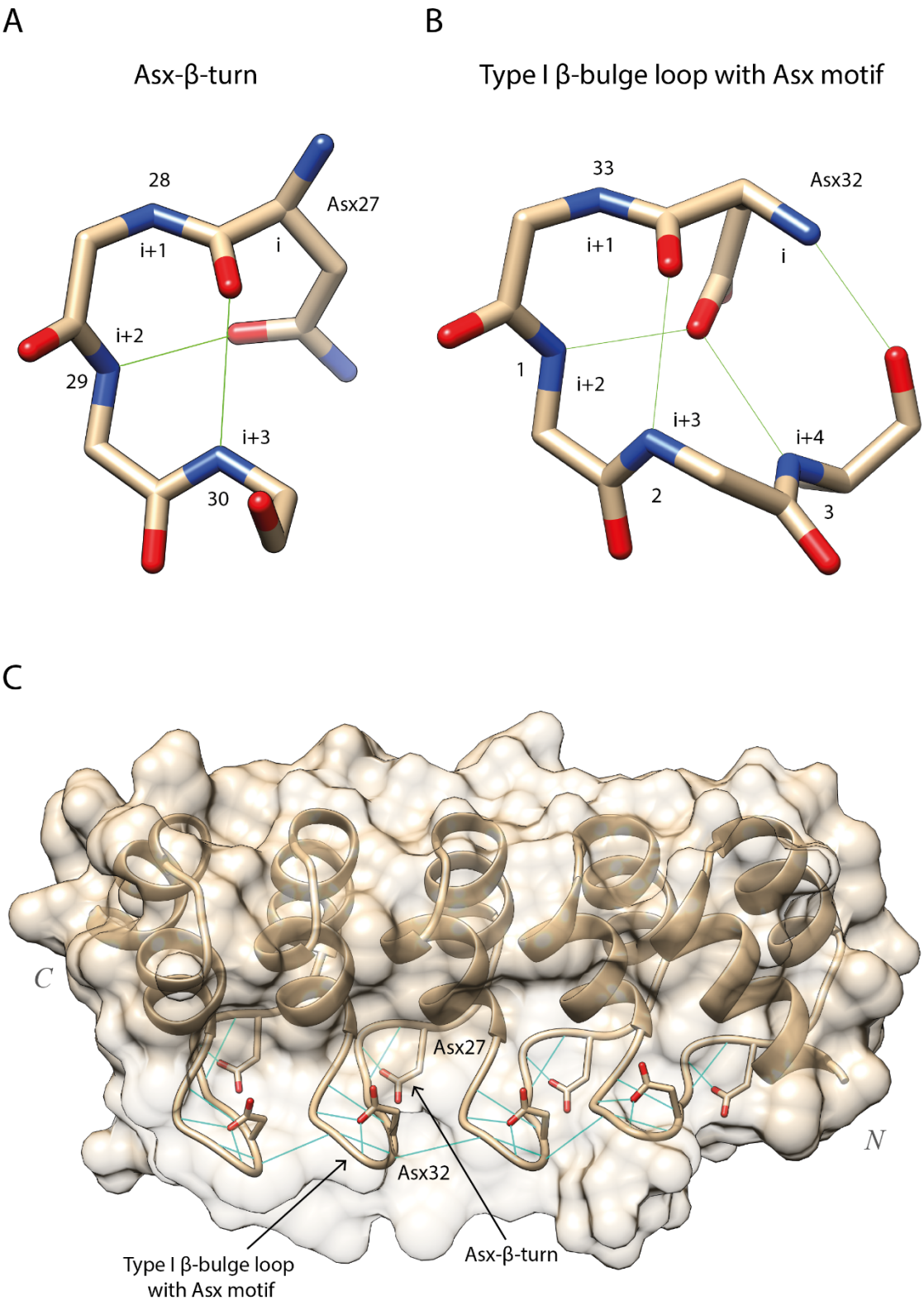

Figure 7. Hydrogen bonding patterns of the two Asx motifs found in the ANK and their location within the ARD. Only repeats with either Asn or Asp at these positions will present this hydrogen bonding pattern. (A) Asx-β-turn at positions 27-30. Conserved Asx, i.e., Asp/Asn, side chain at position i = 27 forms extra hydrogen bond with backbone N at position i + 2; (B) Type 1 β-bulge loop with Asx motif at positions 32-3. Conserved Asx side chain at domain position i = 32 forms two hydrogen bonds with backbone N of residues i + 2 and i + 4. The rest of the hydrogen bonds originate from the backbone of the residues and are not specific of Asx motifs. PDB: 5MA3 (14); (C) DARPin-8.4 (Barandun J, Schroeder T, Mittl PRE,

Grutter MG). Light blue lines represent the hydrogen bonds that determine these secondary structure motifs. The conservation of these Asx residues at positions 27 and 32, together with their depletion in missense variants (Fig 9) and the hydrogen bonding network they facilitate, suggests that these Asx motifs are one of the most structurally important components of the ankyrin repeat domain structure. Figure obtained with UCSF Chimera (7).

Relative solvent accessibility and surface classification

The surface of the ankyrin repeat domain has previously been divided into two faces: concave (positions 32-12) and convex (positions 13-31) (Fig 8A, B) (47). Positions with high RSA (RSA ≈ 50%), such as 1, 12 and 33 are found near positions 13 and 32. Due to their high solvent accessibility, these positions were used to define ridges at the limits of the concave and convex surface. However, our analysis of all available structures also showed positions 23 and 25 to have a high RSA (Fig 6C). In addition, in the same fashion as positions 1, 33 and 12, positions 23 and 25 from different repeats form a ridge on the domain structure. This ridge suggests the definition of a third surface of the domain or basal surface as shown in Figure 8C and enabled the classification of all positions that were not buried into one of the three defined surfaces, (Table 1). This classification is shown in Figure 8C for an ARD containing 12 repeats (48).

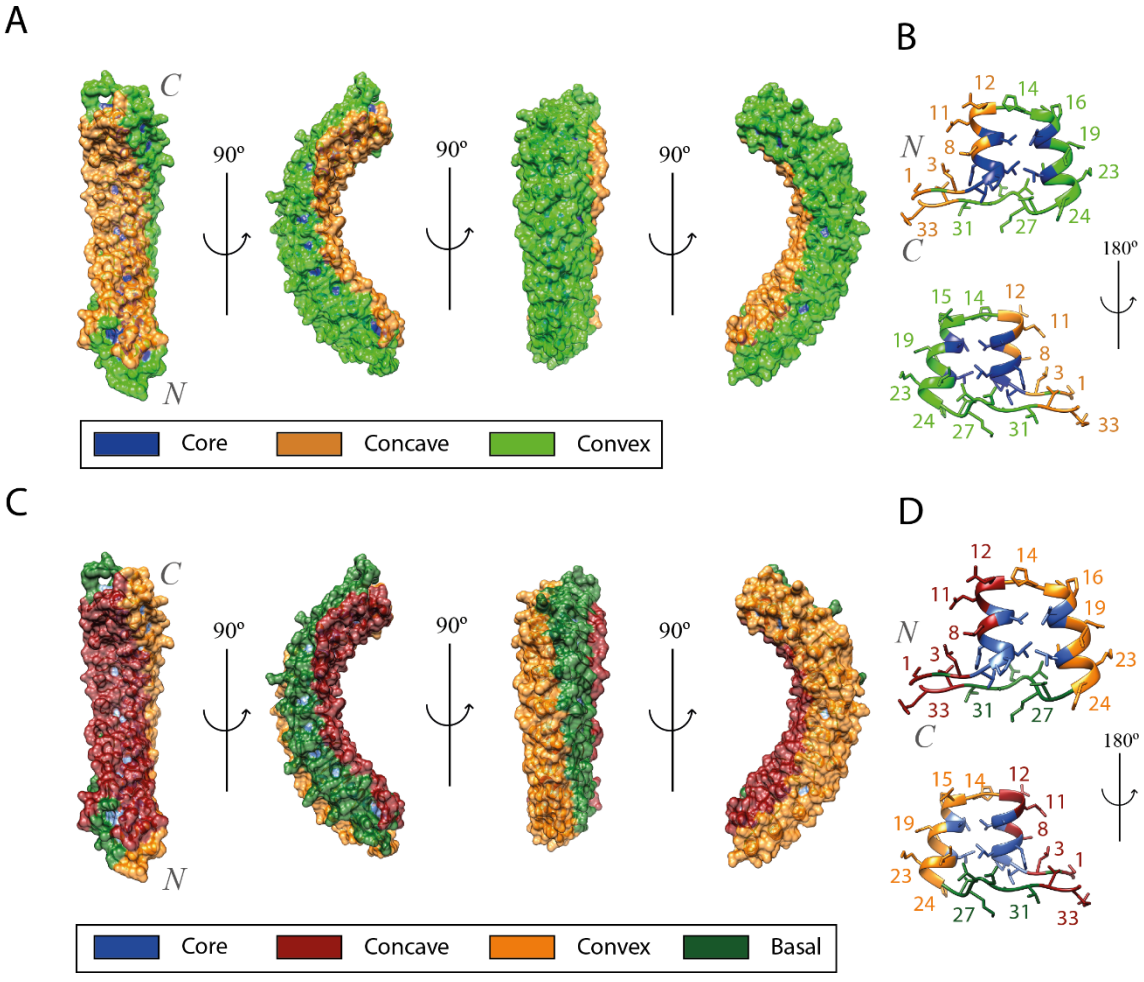

Figure 8. Comparison of the original definition of the ARD surfaces (A, B) with the new definitions derived from the results of this study (C, D). All panels refer to the D34 region of ANK1 ARD, PDB accession: 1N11 (48). This structure shows 12 out of the 23 ARs found on this ARD; (A) Surface of an ARD. Residues conforming the concave surface are coloured in orange, residues on the convex surface in green and buried residues in blue; (B) Surface of an individual repeat. The first α-helix and the β-turn region form the concave surface, whereas the second helix and the loop form the convex one; (C) Residues conforming the concave

surface are coloured in dark red; residues on the convex surface in orange; the basal surface is coloured on dark green and buried residues in blue. Figure obtained with UCSF Chimera (7).

Some blue-coloured regions can be observed on the ARD surface on Figure 8C. These are the side chains of buried residues within the motif dominated by Thr4 and His7. The correct classification of the positions in the ANK as either buried or any of the defined surfaces is critical to calculate accurate enrichment scores in missense variants and protein-protein interactions on a surface basis later in the analysis.

Table 1: Classification of the 33 positions within the ANK in the different surfaces.

Surface	Consensus residue positions
Core	4, 5, 6, 7, 9, 10, 17, 18, 21
Concave	1, 2, 3, 8, 11, 12, 32, 33
Convex	13, 14, 15, 16, 19, 20, 22, 23, 24
Basal	25, 26, 27, 28, 29, 30, 31

Missense variants enrichment analysis

21,338 missense variants from 1,435 human ankyrin repeat sequences were used to calculate column-specific missense enrichment scores (MES). The MES measures how enriched in missense variants an alignment column is compared to the average of the other columns in the alignment (18). The 33 columns of the motif were classified into four categories according to their normalised Shenkin divergence score ($N_{Shenkin}$) and MES. Columns with $0 \leq N_{Shenkin} \leq 25$ and MES < 0 were classified as conserved and missense depleted (CMD), whereas columns satisfying $0 \leq N_{Shenkin} \leq 25$ and MES > 0 were called conserved and missense enriched (CME). We also classified those columns with $75 \leq N_{Shenkin} \leq 100$ as unconserved and either missense depleted (UMD) if MES < 0 or enriched if MES > 0 (UME). $N_{Shenkin}$ ranges from 0 for the most conserved (Position 9) to 100 for the most divergent (Position 3) column within the ANK alignment. As a consequence, positions with $N_{Shenkin}$ < 25 will be those with divergence scores between the minimum and ¼ of the maximum score and positions with $N_{Shenkin}$ > 75will be those with divergence scores on the fourth quartile of the range. Figure 9A shows the enrichment in human population missense variants per position in the ANK relative to their Shenkin divergence score. Positions that are depleted in missense variants relative to the rest of positions within the ANK are the most interesting and are likely to be functionally important. Depletion in missense variation directly results from evolutionary constraint within the human population and is therefore indicative of functional and/or structural relevance (18).

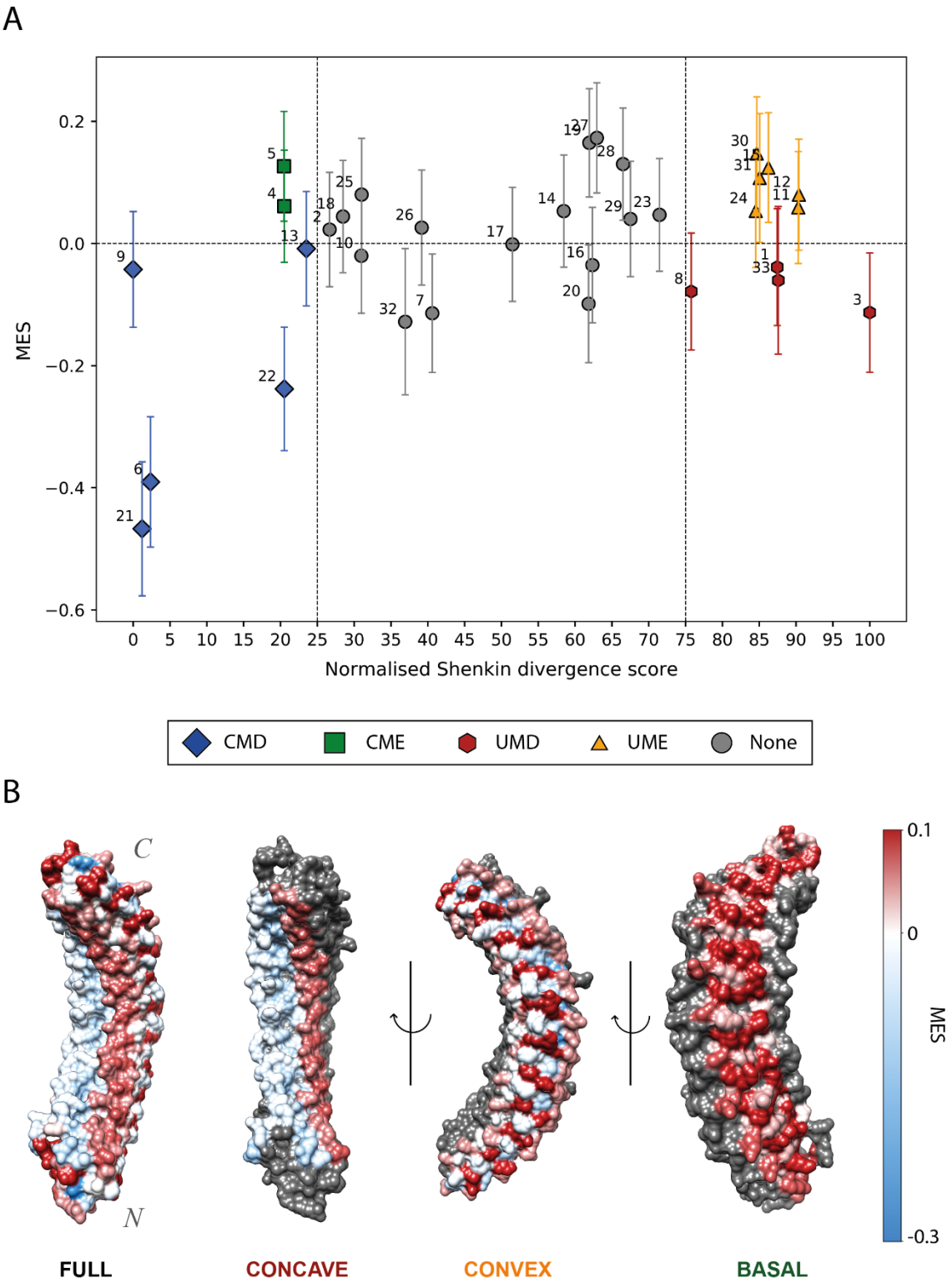

Figure 9. (A) Relative Missense Enrichment Score (MES) against normalised Shenkin divergence score for the 33 positions of the domain. Blue diamonds: CMD positions (6, 9, 13, 21, 22); Green squares: CME positions (4, 5), UMDs are coloured in red hexagons (1, 3, 8, 33) and UMEs in orange triangles (11, 12, 15, 23, 24, 30, 31). Error bars represent 95% CI of the MES, i.e., ln (OR). Error bars for the Shenkin score are not shown as the uncertainty associated with it is negligible. Positions coloured in grey circles are classed as "None", for they do not meet our divergence score thresholds; (B) D34 region of ANK1 ARD, PDB accession: 1N11 (48) This structure shows 12 out of the 23 ARs found on this ARD. Residues are

coloured according to the missense enrichment score of the alignment column they align to in the MSA. The colour scale goes from blue (missense-depleted) to red (missense-enriched) going through white (neutral). From left to right, the full domain, then concave, convex and basal surface are coloured. On each of the last three representations, only one surface is coloured. Residues not belonging in said surface are coloured in grey. Overall, the concave surface is coloured in a light blue colour (except positions 11 and 12), indicating its depletion in missense variants, relative to the other positions within the ANK. Figure obtained with UCSF Chimera (7).

The relationship between residue solvent accessibility and enrichment in missense variants was examined. As expected, on average, buried residues (RSA ≤ 5%) were depleted in missense variants relative to residues present on the surface ($MES = -0.10, p = 1.9 \times 10^{-7}$). Furthermore, residues present on the concave surface of the ankyrin repeat domain were significantly depleted in missense variants relative to the other surfaces, ($MES = -0.08, p = 4.4 \times 10^{-4}$). The convex surface was neither enriched nor depleted, whereas the basal surface was significantly enriched in missense variants: ($MES = 0.09, p = 6.2 \times 10^{-6}$). Moreover, the basal surface is significantly enriched in missense variation relative to the convex one ($MES = 0.08, p = 8.8 \times 10^{-4}$). These results can be observed in structure in Figure 9B and are further discussed in *"Different surfaces of the ARD"* below.

Ankyrin repeat contact maps and enrichment

In this work, contacts across all known ankyrin repeat structures were considered instead of just a single repeat or domain. This allowed a comprehensive contact map for the repeat motif to be calculated as well as enrichment scores for each residue's contacts within the repeat, thus highlighting the most structurally important positions.

Intra-repeat contacts

Figure 10A shows the symmetric contact matrix that defines the ankyrin repeat motif. Contacts between residues within 2-5 amino acids of each other are around the diagonal. Most other contacts are between the residues along first and the second α-helices, from positions 5-11 and 15-23, respectively or contacts between residues close in sequence within the loops. This pattern of contacts is typical of helical or turn secondary structures. Accordingly, we focused on contacts most relevant to the ANK fold, i.e., helix-helix contacts, by filtering out contacts between positions within ≤ 6 residues of each other. Figure 10B shows the enrichment in these intra-repeat contacts for each position within the ANK. CMD positions are among the most enriched in intra-repeat interactions which suggests an important role in ankyrin repeat packing and may explain their conservation across homologs and depletion in missense variants within the human population.

401

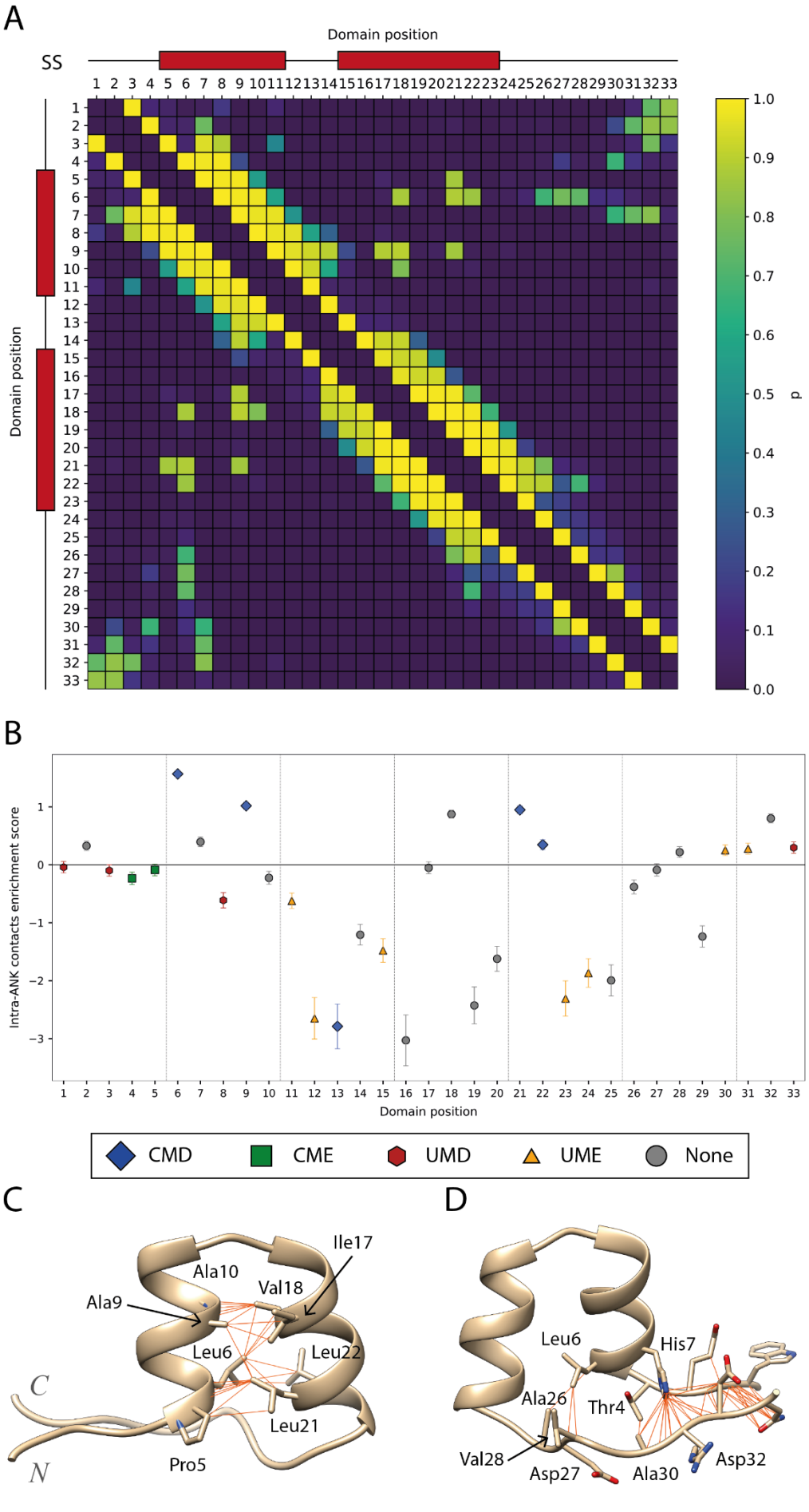

Figure 10. (A) Contact map for intra-repeat residue-residue interactions. Cells are coloured according to the probability of observing contact between two positions with the viridis colour palette. Red boxes above axis indicate the location of the secondary structure (SS) elements, α-helices, in the motif; (B) Intra-repeat contacts enrichment plot. Error bars indicate 95% CI of the enrichment score, i.e., ln (OR). Data points are coloured according to their missense enrichment and residue conservation classification (Fig 9); (C) Cluster of intra-repeat contacts between the first and second helices. Residues 5, 6, 9 and 10 in the first helix interact with residues 17, 18, 21 and 22 by forming hydrophobic interactions. These positions are all buried and conserved; (D) Cluster of intra-repeat contacts between the start and end residues of an AR. These interactions are not as specific as the ones in the first cluster and they include diverse positions such as 1, 3, 31 or 33. These are the most frequently observed contacts across all structure displayed in an example repeat. PDB: 5MA3 (14). Figure obtained with UCSF Chimera (7).

Protein-substrate interaction enrichment

Figure 11A shows the enrichment in Protein-Protein interactions (PPIs) per position in the ANK. Out of the 176 protein structures that satisfied our quality thresholds, as described in Methods, 63 include protein substrates. These represent the interaction between 35 different ARDs and their substrates, accounting for a total of 142 repeats. All the positions that are found on the concave surface are enriched in PPIs. His7 is highly conserved and even though it is buried, part of its side chain is accessible to the concave surface. This way, position 7 interacts with the substrate and appears enriched in the analysis. Positions 13 and 14 define the beginning of the convex surface, at the loop between the two helices. These positions, despite not forming part of the concave surface, are very close to it. Thus, they are enriched in PPIs, though not as significantly as those positions on the concave surface.

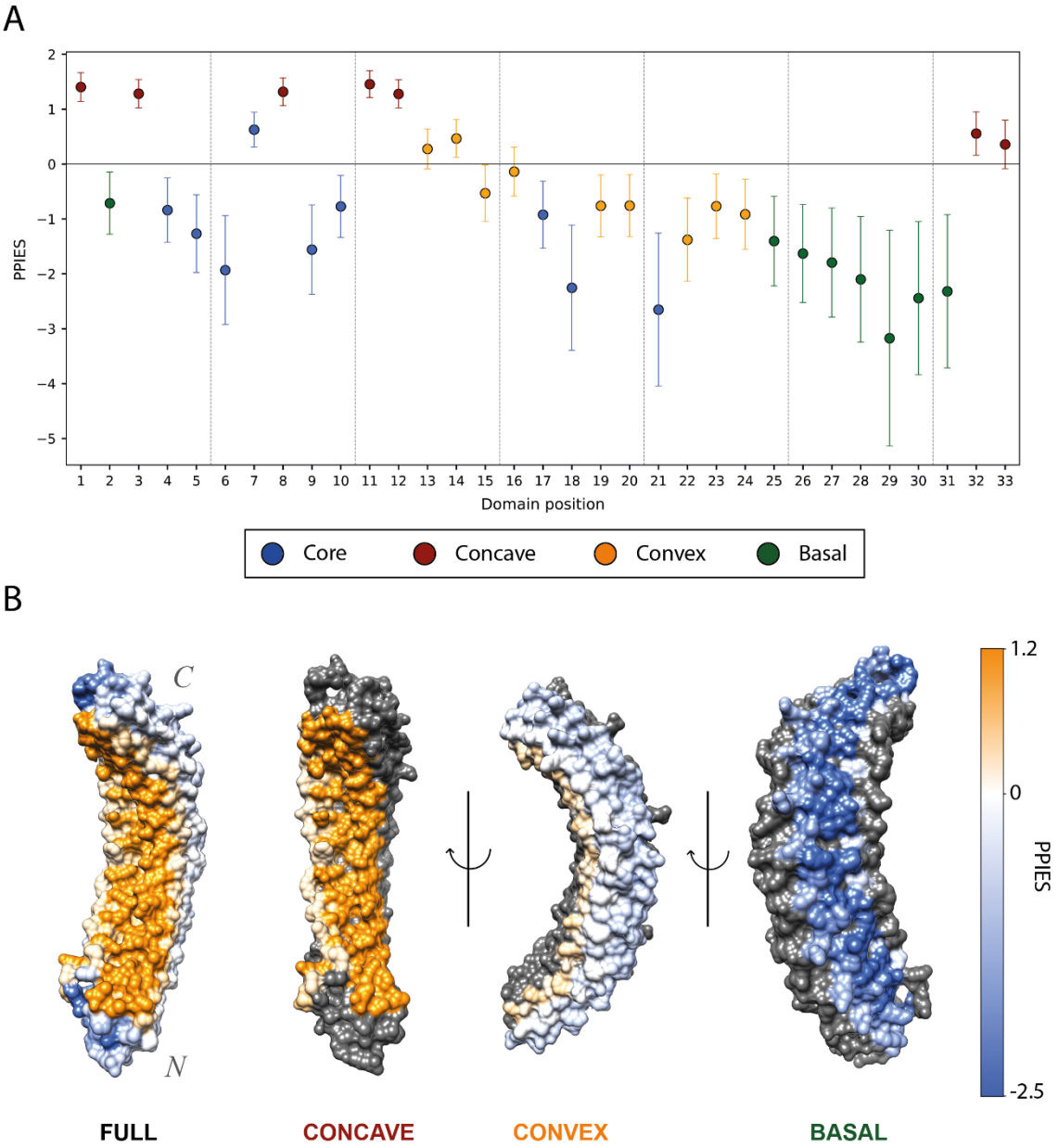

Figure 11. (A) Protein-substrate interactions enrichment plot. Error bars indicate 95% CI of the protein-protein interactions enrichment score (PPIES), i.e., ln (OR). Data points are coloured according to their surface classification (Table 1); (B) D34 region of ANK1 ARD, PDB accession: 1N11 (48). This structure shows 12 out of the 23 ARs found on this ARD. Residues are coloured according to the PPIES of the alignment column they align to in the MSA. The colour scale goes from blue (depleted in PPIs) to orange (enriched in PPIs) going through white (neutral). From left to right, the whole domain, then the concave, convex and basal surface are coloured. On each of the last three representations, only one surface is coloured. Residues not belonging in that surface are coloured in grey. Overall, the concave surface is coloured in a strong orange colour, indicating its importance in protein binding, whereas the basal one presents a dark blue colour, indicative of its overall depletion in PPIs. Figure obtained with UCSF Chimera (7).

We also compared the enrichment in PPIs between different surfaces. As expected, buried residues were significantly depleted on average relative to surface residues, ($PPIES = -1.02, p < 10^{-16}$). Compared to residues belonging to other surfaces, concave residues are highly enriched in PPIs: ($PPIES = 1.86, p < 10^{-16}$). Conversely, convex and basal residues are both depleted in PPIs relative to residues present in the other surfaces: ($PPIES = -0.79, p <$

10^{-16}) and ($PPIES = -2.19, p < 10^{-16}$), respectively. In addition, the direct comparison between basal and convex showed that there is a significant difference regarding their involvement in substrate binding. Residues present on the convex surface were in average enriched in PPIs relative those in the basal surface: ($PPIES = 1.31, p = 5.40 \times 10^{-15}$). All these differences in enrichment in PPIs between different surfaces can be observed in Figure 11B, which shows the different surfaces of an ARD, where residues are coloured according to the PPIES of the column they align to. These results agree with (8) and show the prevalence of the concave surface in substrate binding. They also illustrate the rare, though existing, convex binding as well as the practically null contribution of the basal surface to substrate binding.

Different surfaces of the ARD

In this work, we define the binding mode of an ARD as given by the number of repeats and residues that bind the substrate, as well as the surface the latter belong to. These modes can either be absolute or combined/mixed. In the former, one surface dominates the binding, whereas in the latter, a combination of different surfaces accounts for most of the substrate binding residues. For this part of the analysis, only those proteins with a minimum of two repeats and four residues binding the substrate were considered. Of the remaining 25 proteins, 21 (84%) presented a concave binding mode, only one (4%) presented a convex binding mode and none presented a basal mode. The other three (12%) proteins presented a mixed binding mode, where concave, convex, basal and even buried residues participate in the substrate binding.

These binding surfaces present different patterns of enrichment in variation as well as protein-substrate interactions. The concave surface is significantly depleted in missense variants and enriched in PPIs, whereas the basal is the complete opposite and is enriched in missense variants and depleted in PPIs. These results confirm the dominance of the concave binding mode. In addition, we have observed that ARDs can also present a convex binding mode (49), whereas no basal binding mode was observed in the dataset. The differential importance in substrate binding seems to influence the distribution of missense variants within the motif.

Conserved and missense depleted positions

Positions 6, 9, 21 and 22 were found to be highly conserved and depleted in missense variants relative to the other motif positions (CMD). These positions are mostly buried, and present hydrophobic residues. CMDs are enriched in intra-repeat contacts. This population-level constraint agrees with the amino acid conservation and is proof of the structural relevance of these residues. Positions 7 and 32 are not as conserved as the residues we have classified as CMD; however, they are significantly depleted in missense variants as well. These two residues are structurally relevant due to the hydrogen bonding networks they create, as can be seen in Figure 2C, E.

Unconserved and missense depleted positions

It is known that the concave surface allows high sequence variability, in order to accommodate the diversity of protein substrates that ankyrins bind (8). Positions 1, 3, 8 and 33 are amongst the most diverse positions within the ANK, though at the same time depleted in missense variants in the human population. These positions are enriched in PPIs ($PPIES = 3.6, p < 10^{-16}$) and constitute most of the concave surface of the ARD. Missense depletion at these sites show that they are constrained at a population level, thus confirming the functional importance of these residues.

Figure 12 illustrates how the ARDs of homolog pairs with gene names ANKRA2/RFXANK and TNKS1/TNKS2 bind their protein substrates. ANKRA2 and RFXANK are human proteins that are involved in the regulation of transcription by RNA polymerase II. Both ankyrin repeat domains present five repeat units. The domains are very similar in sequence, including UMD and unconserved positions 11 and 12, which do not vary across these proteins' homologs. Multiple structures have been solved portraying the interaction between these ARDs and more than five different protein substrates. All the substrates present the shared binding motif PXLPX[I/L] (50) (51). A similar pattern can be observed with TNKS1 and TNKS2 and the substrates they bind, which share the tankyrase binding motif RXXPDG (52).

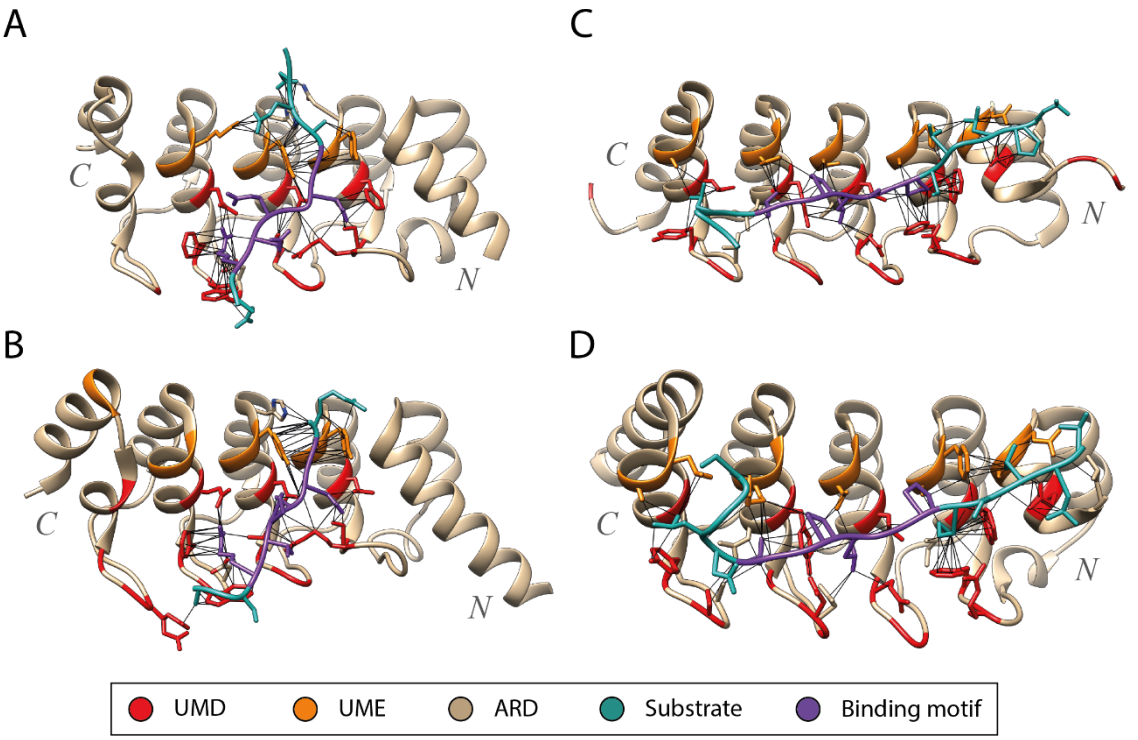

Figure 12. ARDs in complex with substrates. (A) RFXANK and RFX5 (PDB ID: 3V30) (50); (B) ANRA2 and HDAC4 (3V31) RN1262; (C) TNKS2 and ARPIN (4Z68) (53); (D) TNKS1 and USP25 (5GP7) (54). UMD positions (red) and UMEs 11, 12 (orange) are conserved across proteins that bind similar substrates (dark cyan). For example, these positions are conserved across TNKS2 and TNKS1, which are known to bind substrates with the motif RXXPDG (purple). Similarly, RFXANK and ANRA2, bind substrates with the motif PXLPX[I/L] (purple). Figure obtained with UCSF Chimera (7).

These examples show how ankyrin domains that present similar concave surfaces, determined by their UMD positions (1, 3, 8 and 33) bind similar protein substrates, or at least, substrates that share a binding motif. At the same time, it seems that all substrates binding these domains share a binding motif. These findings further support the hypothesis presented by MacGowan, Madeira (18), which states that UMDs are determinant for substrate binding specificity.

Conclusions

The multiple sequence alignment of homologues and the aggregation of genetic variants, or other features, over alignment columns, as described in MacGowan, Madeira (18), can provide insight at the residue level on the evolutionary constraint acting on functional domains as well as highlight structural or functionally relevant residues in protein domains. Overall, a clear variation distribution pattern can be observed within the ankyrin repeat motif. There are five positions that

are conserved and depleted in missense variation due to their structural importance, e.g., enrichment in intra-repeat contacts. Four other positions are highly variable within the family and overall depleted in missense variants, for they are the key for a specific and successful substrate binding.

In this study, we use 7,407 ankyrin repeat sequences, 21,338 human missense variants and 160 3D structures to study the distribution of missense variants within the ankyrin repeat motif and explain the observed patterns with structural data. The general conclusions are as follows.

1. Two of the turns found on the secondary structure of the ANK, positions 28-29 and 33-1, are Asx motifs. Positions 27 and 32 present conserved Asx.
2. The surface of the ARD can be divided in three different surfaces using the RSA of the repeat positions.
3. Positions that are conserved and depleted in missense variants (CMD) are significantly enriched in intra-repeat contacts ($OR = 2.8, p \approx 0$) and are key for the correct packing of the motif as well as the domain.
4. Positions that are unconserved yet depleted in missense variants (UMD) are heavily enriched in protein-protein interactions ($OR = 3.6, p < 10^{-16}$) and might be responsible for substrate binding specificity in the motif.
5. The concave surface of the ARD is significantly enriched in PPIs ($PPIES = 1.86, p < 10^{-16}$) and consequently depleted in missense variation ($MES = -0.08, p = 4.4 \times 10^{-4}$) whereas the other two surfaces are less constrained in line with their reduced importance in substrate binding.

Acknowledgements

We thank Drs Marek Gierlinski, Matthew Parker and Jim Procter for their insightful suggestions during this research. We also thank Prof. Ulrich Zachariae for critical reading of the manuscript and the IT service of the University of Dundee for their support of our HPC infrastructure. This work was supported by Biotechnology and Biological Sciences Research Council Grants (BB/J019364/1 and BB/R014752/1) and Wellcome Trust Biomedical Resources Grant (101651/Z/13/Z).

References

1. Bork P. Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile modules that cross phyla horizontally? Proteins. 1993;17(4):363-74.
2. Andrade MA, Perez-Iratxeta C, Ponting CP. Protein repeats: structures, functions, and evolution. J Struct Biol. 2001;134(2-3):117-31.
3. Sedgwick SG, Smerdon SJ. The ankyrin repeat: a diversity of interactions on a common structural framework. Trends Biochem Sci. 1999;24(8):311-6.
4. Forrer P, Stumpp MT, Binz HK, Pluckthun A. A novel strategy to design binding molecules harnessing the modular nature of repeat proteins. FEBS Lett. 2003;539(1-3):2-6.
5. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188-90.
6. Krzywda S, Brzozowski AM, Higashitsuji H, Fujita J, Welchman R, Dawson S, et al. The crystal structure of gankyrin, an oncoprotein found in complexes with cyclin-dependent kinase 4, a 19 S proteasomal ATPase regulator, and the tumor suppressors Rb and p53. J Biol Chem. 2004;279(2):1541-5.
7. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605-12.

560 8. Mosavi LK, Minor DL, Jr., Peng ZY. Consensus-derived structural determinants of the
561 ankyrin repeat motif. Proc Natl Acad Sci U S A. 2002;99(25):16029-34.
562 9. Islam Z, Nagampalli RSK, Fatima MT, Ashraf GM. New paradigm in ankyrin repeats:
563 Beyond protein-protein interaction module. Int J Biol Macromol. 2018;109:1164-73.
564 10. Stumpp MT, Binz HK, Amstutz P. DARPins: a new generation of protein therapeutics.
565 Drug Discov Today. 2008;13(15-16):695-701.
566 11. Main ER, Jackson SE, Regan L. The folding and design of repeat proteins: reaching a
567 consensus. Curr Opin Struct Biol. 2003;13(4):482-9.
568 12. Kohl A, Binz HK, Forrer P, Stumpp MT, Pluckthun A, Grutter MG. Designed to be stable:
569 crystal structure of a consensus ankyrin repeat protein. Proc Natl Acad Sci U S A.
570 2003;100(4):1700-5.
571 13. Li J, Mahajan A, Tsai MD. Ankyrin repeat: a unique motif mediating protein-protein
572 interactions. Biochemistry. 2006;45(51):15168-78.
573 14. Hansen S, Stuber JC, Ernst P, Koch A, Bojar D, Batyuk A, et al. Design and applications
574 of a clamp for Green Fluorescent Protein with picomolar affinity. Sci Rep. 2017;7(1):16292.
575 15. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to
576 functional variation and the interpretation of personal genomes. PLoS Genet.
577 2013;9(8):e1003709.
578 16. Gussow AB, Petrovski S, Wang Q, Allen AS, Goldstein DB. The intolerance to functional
579 genetic variation of protein domains predicts the localization of pathogenic mutations within
580 genes. Genome Biol. 2016;17:9.
581 17. Sivley RM, Dou X, Meiler J, Bush WS, Capra JA. Comprehensive Analysis of Constraint
582 on the Spatial Distribution of Missense Variants in Human Protein Structures. Am J Hum Genet.
583 2018;102(3):415-26.
584 18. MacGowan SA, Madeira F, Britto-Borges T, Schmittner MS, Cole C, Barton GJ. Human
585 Missense Variation is Constrained by Domain Structure and Highlights Functional and
586 Pathogenic Residues. bioRxiv. 2017:127050.
587 19. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The
588 mutational constraint spectrum quantified from variation in 141,456 humans. bioRxiv.
589 2020:531210.
590 20. UniProt C. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res.
591 2019;47(D1):D506-D15.
592 21. Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture
593 research tool: identification of signaling domains. Proc Natl Acad Sci U S A. 1998;95(11):5857-
594 64.
595 22. Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, et al. New and continuing
596 developments at PROSITE. Nucleic Acids Res. 2013;41(Database issue):D344-7.
597 23. Attwood TK. The PRINTS database: a resource for identification of protein families.
598 Brief Bioinform. 2002;3(3):252-63.
599 24. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein
600 families database in 2019. Nucleic Acids Res. 2019;47(D1):D427-D32.
601 25. Sievers F, Higgins DG. Clustal Omega, accurate alignment of very large numbers of
602 sequences. Methods Mol Biol. 2014;1079:105-16.
603 26. Eddy SR. Multiple alignment using hidden Markov models. Proc Int Conf Intell Syst Mol
604 Biol. 1995;3:114-20.
605 27. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate
606 multiple sequence alignment. J Mol Biol. 2000;302(1):205-17.
607 28. Barton GJ. The AMPS package for multiple protein sequence alignment. Methods Mol
608 Biol. 1994;25:327-47.
609 29. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high
610 throughput. Nucleic Acids Res. 2004;32(5):1792-7.

30. Russell RB, Barton GJ. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins. 1992;14(2):309-23.
31. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. Multiple sequence alignment with Clustal X. Trends Biochem Sci. 1998;23(10):403-5.
32. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25(9):1189-91.
33. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Biol. 2003;10(12):980.
34. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. 2007;35(Database issue):D301-3.
35. wwPDBconsortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. 2019;47(D1):D520-D8.
36. Velankar S, Dana JM, Jacobsen J, van Ginkel G, Gane PJ, Luo J, et al. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res. 2013;41(Database issue):D483-9.
37. MacGowan SA, Madeira F, Britto-Borges T, Warowny M, Drozdetskiy A, Procter JB, et al. The Dundee Resource for Sequence Analysis and Structure Prediction. Protein Sci. 2020;29(1):277-97.
38. Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wahlby A, Jones TA. The Uppsala Electron-Density Server. Acta Crystallogr D Biol Crystallogr. 2004;60(Pt 12 Pt 1):2240-9.
39. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577-637.
40. Shenkin PS, Erman B, Mastrandrea LD. Information-theoretical entropy as a measure of sequence variability. Proteins. 1991;11(4):297-313.
41. Szumilas M. Explaining odds ratios. J Can Acad Child Adolesc Psychiatry. 2010;19(3):227-9.
42. Jubb HC, Higueruelo AP, Ochoa-Montano B, Pitt WR, Ascher DB, Blundell TL. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. J Mol Biol. 2017;429(3):365-71.
43. Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. Maximum allowed solvent accessibilites of residues in proteins. Plos One. 2013;8(11):e80635.
44. Miller S, Janin J, Lesk AM, Chothia C. Interior and surface of monomeric proteins. J Mol Biol. 1987;196(3):641-56.
45. de Brevern AG. Extension of the classical classification of beta-turns. Sci Rep. 2016;6:33191.
46. Wan WY, Milner-White EJ. A natural grouping of motifs with an aspartate or asparagine residue forming two hydrogen bonds to residues ahead in sequence: their occurrence at alpha-helical N termini and in other situations. J Mol Biol. 1999;286(5):1633-49.
47. Wang C, Wei Z, Chen K, Ye F, Yu C, Bennett V, et al. Structural basis of diverse membrane target recognitions by ankyrins. Elife. 2014;3.
48. Michaely P, Tomchick DR, Machius M, Anderson RG. Crystal structure of a 12 ANK repeat stack from human ankyrinR. EMBO J. 2002;21(23):6387-96.
49. Hesketh GG, Perez-Dorado I, Jackson LP, Wartosch L, Schafer IB, Gray SR, et al. VARP is recruited on to endosomes by direct interaction with retromer, where together they function in export to the cell surface. Dev Cell. 2014;29(5):591-606.
50. Xu C, Jin J, Bian C, Lam R, Tian R, Weist R, et al. Sequence-specific recognition of a PxLPxI/L motif by an ankyrin repeat tumbler lock. Sci Signal. 2012;5(226):ra39.
51. Nie J, Xu C, Jin J, Aka JA, Tempel W, Nguyen V, et al. Ankyrin repeats of ANKRA2 recognize a PxLPxL motif on the 3M syndrome protein CCDC8. Structure. 2015;23(4):700-12.

662 52. Sbodio JI, Chi NW. Identification of a tankyrase-binding motif shared by IRAP, TAB182,
663 and human TRF1 but not mouse TRF1. NuMA contains this RXXPDG motif and is a novel
664 tankyrase partner. J Biol Chem. 2002;277(35):31887-92.
665 53. Fetics S, Thureau A, Campanacci V, Aumont-Nicaise M, Dang I, Gautreau A, et al.
666 Hybrid Structural Analysis of the Arp2/3 Regulator Arpin Identifies Its Acidic Tail as a Primary
667 Binding Epitope. Structure. 2016;24(2):252-60.
668 54. Xu W, Lau YH, Fischer G, Tan YS, Chattopadhyay A, de la Roche M, et al. Macrocyclized
669 Extended Peptides: Inhibiting the Substrate-Recognition Domain of Tankyrase. J Am Chem Soc.
670 2017;139(6):2245-56.

671