Received: 24 February 2022 | Accepted: 6 April 2022

DOI: 10.1002/pro.4317

ACCELERATED COMMUNICATION

THE PROTEIN SOCIETY WILEY

# The return of the rings: Evolutionary convergence of aromatic residues in the intrinsically disordered regions of RNA-binding proteins for liquid–liquid phase separation

**Wen-Lin Ho**[1] | **Jie-rong Huang**[1,2,3]

[1]Institute of Biochemistry and Molecular Biology, National Yang Ming Chiao Tung University, Taipei, Taiwan

[2]Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei, Taiwan

[3]Department of Life Sciences and Institute of Genome Sciences, National Yang Ming Chiao Tung University, Taipei, Taiwan

**Correspondence**
Jie-rong Huang, Institute of Biochemistry and Molecular Biology, National Yang Ming Chiao Tung University,
No. 155 Section 2, Li-Nong Street, Taipei, Taiwan.
Email: jierongh@nycu.edu.tw

**Funding information**
The Ministry of Science and Technology of Taiwan, Grant/Award Numbers: 109-2113-M-010-003, 110-2113-M-A49A-504-MY3

**Review editor:** Nir Ben-Tal

## Abstract

Aromatic residues appeared relatively late in the evolution of protein sequences to stabilize the globular proteins' folding core and are less in the intrinsically disordered regions (IDRs). Recent advances in protein liquid–liquid phase separation (LLPS) studies have also shown that aromatic residues in IDRs often act as "stickers" to promote multivalent interactions in forming higher-order oligomers. To study how general these structure-promoting residues are in IDRs, we compared levels of sequence disorder in RNA binding proteins (RBPs), which are often found to undergo LLPS, and the human proteome. We found that aromatic residues appear more frequently than expected in the IDRs of RBPs and, through multiple sequence alignment analysis, those aromatic residues are often conserved among chordates. Using TDP-43, FUS, and some other well-studied LLPS proteins as examples, the conserved aromatic residues are important to their LLPS-related functions. These analyses suggest that aromatic residues may have contributed twice to evolution: stabilizing structured proteins and assembling biomolecular condensates.

**KEYWORDS**

biomolecular condensates, intrinsically disordered proteins, liquid–liquid phase separation, membraneless organelle, RNA-binding proteins

Amino acids with aromatic rings stabilize protein structure[1] and are accordingly rare in intrinsically disordered regions (IDRs).[2,3] Interestingly, however, aromatic residues in IDRs have recently been identified as being crucial in mediating liquid–liquid phase separation (LLPS).[4–7] Here, through protein sequence analyses, we outline how aromatic residues may have contributed twice to evolution.

How did proteins emerge on Earth? Eck and Dayhoff proposed that early polypeptides were short, with simple compositions, but extended constantly by duplicating their sequences.[8] In their ferredoxin example, new amino acids appeared by mutation, and the addition of cysteine provided sulfide bonding to ferrous sulfide, a catalyst used as a primitive energy source.[8] The more lately incorporated amino acids, including hydrophobic and aromatic residues, enhanced the folding stability of these polypeptides.[1] These polypeptides, or proteins, became the main workhorses of early cells. When lifeforms became more complex with the appearance of eukaryotes

Short statements for broader audience: Aromatic residues appear frequently in the core of folded proteins and are relatively rare in the intrinsically disordered regions (IDRs). However, many IDRs that mediate liquid–liquid phase separation (LLPS) have aromatic residues that are important for their functional assembly. Focusing on RNA-binding proteins (many undergo LLPS), we show that aromatic residues have contributed twice to evolution: first, in stabilizing structured protein, and later, in facilitating LLPS in IDRs.

  

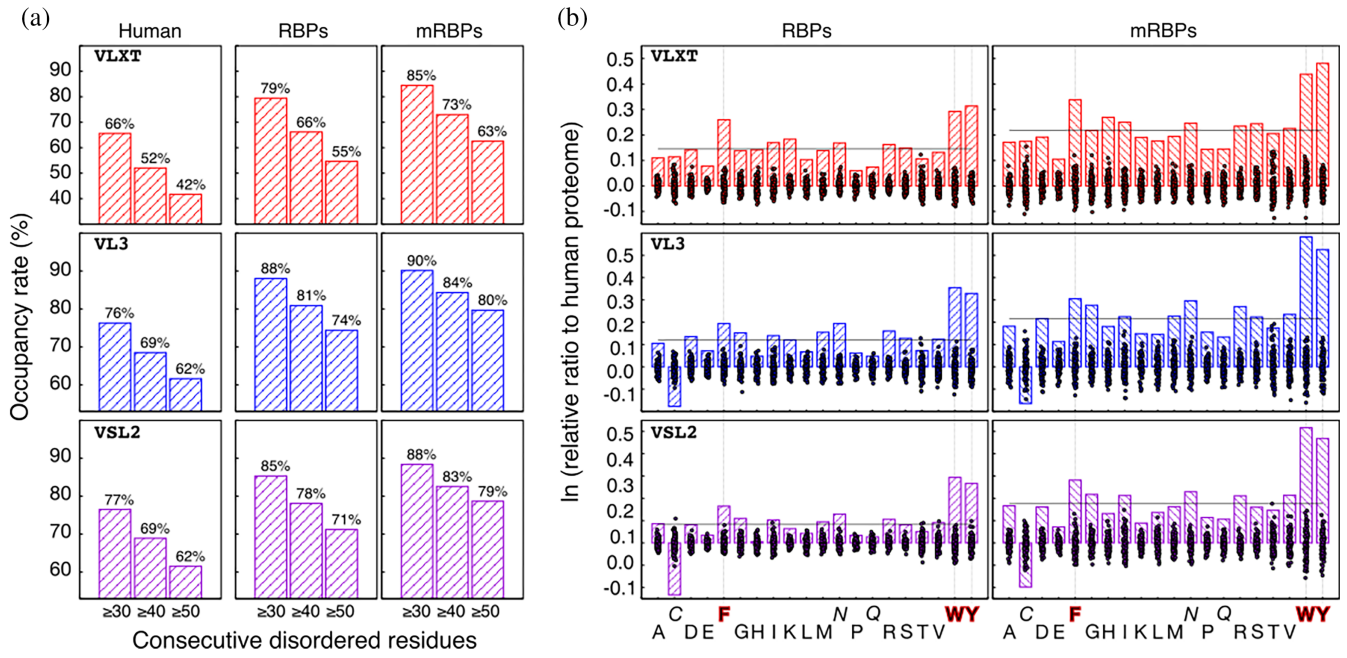


**FIGURE 1** Prevalence of disorder and disorder odds ratios relative to the human proteome by amino acid type for RNA-binding proteins. (a) Proportion of proteins with disordered regions longer than 30, 40, or 50 consecutive residues, as predicted using different algorithms, in the human proteome (left column), RNA-binding proteins (middle column), and mRNA binding proteins (right column). (b) Log-odds ratios relative to the human proteome of being in an intrinsically disordered region for amino acids in RNA binding proteins (RBPs) (left) and mRBPs (right). The dashed lines indicate the average value over all amino acid types. The dots represent the values obtained for randomly selected (negative control) subsamples ($N = 1{,}542$ for RBPs and $N = 689$ for mRBP, the same numbers as considered in the main analysis) of the human proteome

and multicellular organisms, proteins became more efficient and "moonlighted" in different functions.[9] The acquisition of IDRs substantially enlarged their functional repertoire.[10] Among many functional advantages,[11] IDRs' tunable and reversible assembly, known as LLPS, has recently been recognized as the mechanism that drives the formation of membraneless organelles.[12]

The building blocks of proteins also appeared at different stages of evolution. Eck and Dayhoff found, for example, that the earliest four amino acids to appear were Ala, Asp, Ser, and Gly.[8] Based on multiple criteria,[13] the order of appearance of the amino acids in life is now accepted to be: Gly, Ala, Asp, Val, Pro, Ser, Glu, Leu/Thr, Arg, Ile/Gln/Asn, His, Lys, Cys, Phe, Tyr, Met, Trp. Interestingly, as many others have noticed, this suggests that the earliest peptides were disordered because the most primitive amino acids are not structure-promoting.[14–16] These primitive disordered proteins may have initiated life on Earth and acquired aromatic residues later, thereby increasing their folding stability. The initial set of amino acids was implicated again in the evolution of IDRs for various functions, as these disordered regions were "reinvented."[17] Although aromatic residues appear relatively late in protein evolution for stabilizing the folding, they are also present in the molecular recognition features (often phenylalanine) in later evolved IDRs.[18,19] More importantly, recent studies have shown that they are one of the driving forces of LLPS in the IDRs.[4–7] Therefore, we would like to understand how general the appearance of "structure-promoting" amino acids is in unstructured regions.

We are particularly interested in RNA-binding proteins (RBPs) because the evidence is accumulating that their subcellular localization with RNA molecules is mediated by LLPS,[20] and posttranscriptional gene control is dependent on RNA molecules being in precisely the right place at the right time.[21,22] Accordingly, we analyzed a set of 1,542 RBPs, a subgroup of 692 mRNA binding proteins (mRBPs), and other types of RBPs from a census study,[23] in comparison with a set of 20,396 human proteins. We collected protein sequences from the UniProt database[24] and analyzed their level of disorder using the PONDR server's VLXT, VL3, and VSL2 algorithms.[25,26] Although these algorithms predict different results for a single protein, the overall trend in a large-scale analysis is similar (see analysis below and Figure S1). We separated residues into ordered and disordered based on the annotation of these algorithms, and calculated the percentage of proteins with sequences of

consecutive disordered residues longer than 30, 40, and 50 amino acids (Figure 1a; Table S1). Our analysis for the human proteome is similar to a classic study of eukaryotic organisms[27] and agrees with the observation that IDRs are more prevalent in RBPs.[28,29] Our results also show that mRBPs have an even higher proportion of disordered residues (Figure 1a), with ~10 percentage points (pp) more residues than human proteins in consecutive sequences of 20 or more disordered residues, and ~20 pp more residues found in disordered regions longer than 60 amino acids (Table S1), suggesting that the mRBPs tend to have longer IDRs. We also analyzed the subgroups of the other types of RBPs. Our results agree with a recent analysis that mRNA, rRNA, and snRNA binding proteins have a higher portion of disordered regions[30] (Figure S2).

Next, we counted the numbers of each amino-acid type in the predicted disordered or folded domains (Tables S2–S4). For each amino acid, we calculated the log-odds ratio relative to the human proteome of occurring in a disordered region in RBPs or mRBPs (Figure 1b):

$$\text{Log}-\text{odds ratio of presence in IDRs} = ln\left(\frac{\frac{\text{RBP}^{\text{disordered}}}{\text{RBP}^{\text{total}}}}{\frac{\text{Human}^{\text{disordered}}}{\text{Human}^{\text{total}}}}\right) \quad (1)$$

The values obtained are positive for all amino acids (except for cysteine with the VL3 and VSL2 algorithms; Figure 1b), and the average values (dashed lines in Figure 1b) confirm the above conclusion that IDRs are more prevalent in RBPs than in the human proteome. We also repeated the analysis 1,000 times for random selections of 1,542 or 692 proteins sequences (the numbers of RBPs and mRPBP sequences considered) from the human proteome. The averaged log-odds ratios obtained for the random selections are around zero with standard deviations of mostly less than 0.05 and largest deviations no greater than ±0.1 (dots in Figure 1b; Tables S5–S7 and Figure S3). This analysis of random selections indicates that the differences for RBPs and mRBPs (bars in Figure 1b) are significant. We also analyzed the other types of RBPs using the same approach, but the deviation of single amino-acid types is not obvious compared to the random distribution (Figure S2).

The results from the different algorithms for each amino acid are consistent, except for cysteine. VL3 and VSL2 predict that the prevalence of cysteines in IDRs is lower in RBPs and mRBPs than in the human proteome, whereas VLXT predicts the opposite, but similar to the range of results of random selection. Although this may reflect the algorithm's use of different scoring functions, it is also possible that IDRs in human proteins have a higher portion of cysteines than those in RBPs do. Note also that asparagine is slightly more likely to be disordered in RBPs and mRBPs, whereas for glutamine, the difference is similar to those obtained for the randomly selected pools (Figure 1b). Although it has been reported that the IDRs in RBPs are likely to be prion-like, that is, rich in asparagine and glutamine,[7] these two amino acids are not obviously more prevalent than in the human proteome. Among all amino acids, the most considerable differences obtained with these three algorithms are for phenylalanine, tryptophan, and tyrosine, with values much higher than those obtained for random samples of human proteins, particularly for mRNA targeting proteins (Figure 1b). In other words, *structure-promoting* aromatic amino acids are relatively more abundant in *disordered* regions of RBPs than in the human proteome in general.

Functionally important residues are conserved during evolution.[31] Therefore, we aligned orthologue sequences for each RBP and calculated the Jensen–Shannon divergence score of each residue[31] as a measure of sequence conservation (Figure 2; Methods in Supporting Information S1). The level of conservation is lower in the chordate phylum (dark blue lines in Figure 2) than in mammals (light blue lines). The analysis was repeated (gray lines) for vertebrates (subphylum) and tetrapods (superclass). As expected, folded domains are conserved earlier in evolution than disordered regions, as indicated by that the dark blue lines have lower values in the regions underlined with a red bar (sequences of more than 40 consecutive disordered residues). We also calculated the average level of conservation for each amino acid to investigate individual trends. The trends are generally flat for structured domains, indicating early conservation (examples in Figure S4). On the contrary, although increasing trends are observed for most amino acids in the IDRs, the aromatic residues are relatively flat, as observed for residues in structured domains (bottom panels in Figure 2 and Figure S4).

In TDP-43, for example, one of the most extensively studied RBP undergoing LLPS,[32–35] the aromatic residues in its IDRs are highly conserved (Figure 2a; the orange/purple/yellow circles on the Chordata line respectively indicate Phe/Trp/Tyr). The three tryptophans known to be key residues in driving LLPS[4] are conserved in all chordates. The phenylalanines are conserved in chordates but to a lesser extent, in keeping with the fact that they contribute less to LLPS.[4] This interpretation that sequence conservation reflects functional importance is further supported by the fact that the transient α-helical region in TDP-43's IDR (Figure 2a, ~residues 320–340), which is involved in LLPS[32,36,37] is also conserved.

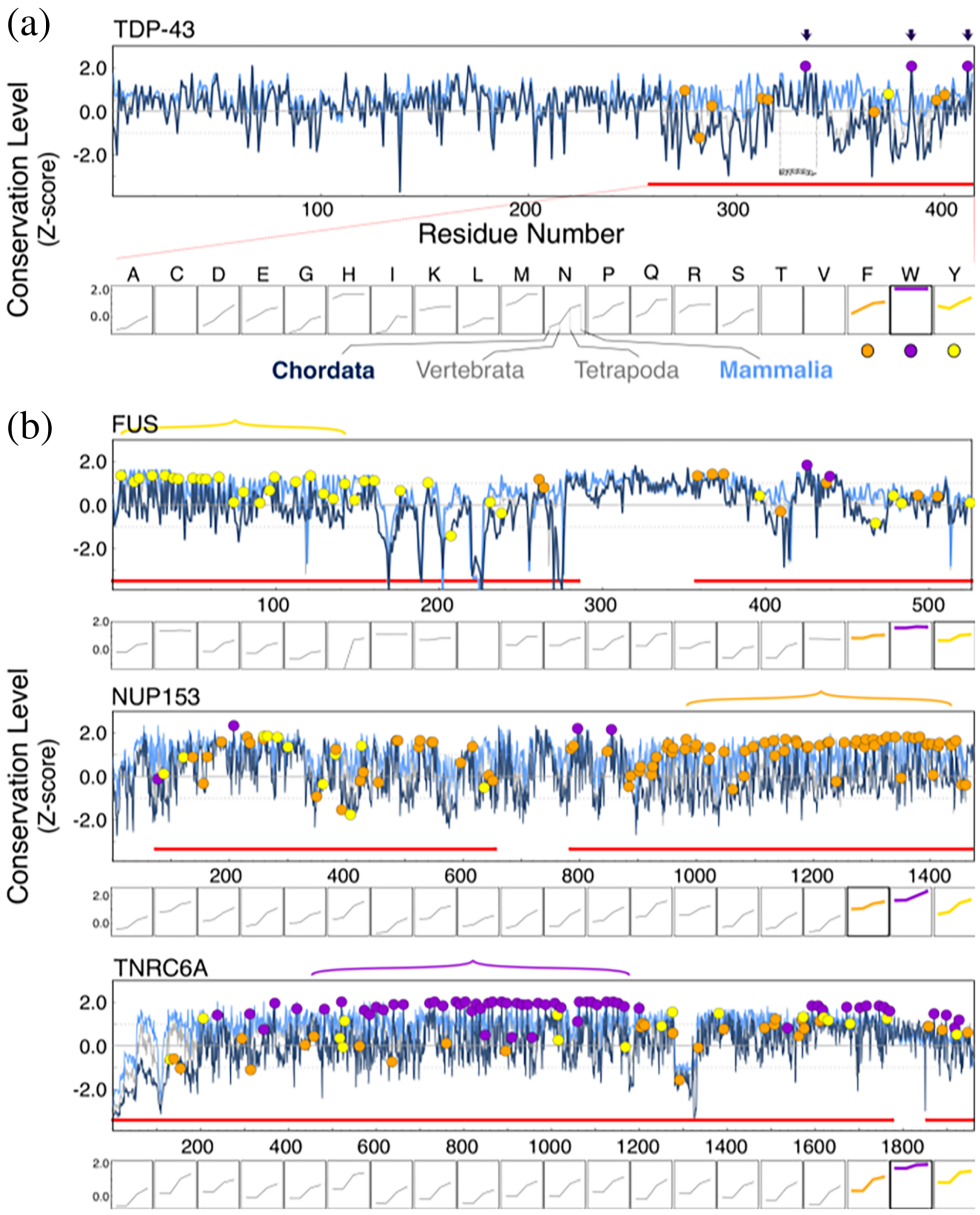


**FIGURE 2** Sequence conservation in example proteins with highly conserved aromatic residues. (a) TDP-43, (b) FUS, NUP153, and TNRC6A. (*Upper panels*) Levels of sequence conservation are quantified by the Jensen–Shannon divergence score[31] and normalized using the Z-score function (the mean of all values toward chordates is 0; the value in the y-axes is the standard deviation, positive values mean more conserved.). Levels of conservation in chordates (dark blue), vertebrates (gray), tetrapods (gray), and mammals (light blue), plotted versus the corresponding residue number in the human sequence. Predicted disordered regions longer than 40 residues are indicated with red bars. Aromatic residues are labeled on the chordate line: Phe (orange), Trp (purple), Tyr (yellow). (a) The three arrows indicate tryptophans experimentally identified as being crucial to liquid–liquid phase separation; the transient α-helical region that also contributes to self-assembly is also labeled. (b) The colored horizontal curly brackets indicate highly conserved aromatic-residue-rich regions. (*Lower panels*) Average levels of conservation as a function of decreasing taxonomic rank for amino-acid types in regions predicted to be disordered (indicated by red bars in the upper panel)

Many other RBPs have highly conserved aromatic residues whose functional importance have been reported (Figure 2b): Tyrosines are known to be involved in the LLPS of the N-terminal domain of the RBP FUS,[5,7,38] and a recent analysis of mammalian FUS proteins found the same trend as observed here.[39] Phe-Gly repeats are a common feature of nucleoporins, which regulate nucleocytoplasmic transport in the nuclear pore complex.[40–42] The nucleoporin NUP153 is also categorized as an RBP because its Phe-Gly repeats mediate mRNA trafficking.[43] Phenylalanines are conserved in NUP153 (Figure 2b) and it is reasonable to suppose that

this is also the case in other nucleoporins with many Phe-Gly repeats. The tryptophan-rich region in TNRC6 interacts with Ago2 to promote the miRNA-induced silencing complex's phase separation[44]; most of these tryptophans are highly conserved in chordates (Figure 2b). Other examples include the LLPS of CPEB[45] and HNRNPD.[46] Although the role of aromatic residues has not been explicitly studied in these proteins, their conservation in chordates hints at a possible role in LLPS related functions (Figure S5). Many RBPs that have not been reported to undergo LLPS also have conserved aromatic patterns, including RBM19, DDX18, KHDRBS1, ABT1, RSL24D1, SMAD5, and TAF9 (Figures S4b and S5). We suggest that these conserved aromatic residues in IDRs may also be involved in LLPS-related functions. However, multiple sequence alignment is limited in identifying the conserved residues in many other IDRs via current algorithms. The analysis above thus underestimates the numbers of RBPs having functionally important aromatic residues. These "unalignable" aromatic residues, however, may also be functionally important. For example, recent studies have shown that aromatic residues' amount and patterning are the keys for hnRNP-A1's LLPS regardless of its orthologs' IDRs cannot be aligned.[47] Moreover, the "spacers" between the stickers are also essential factors in driving LLPS,[47,48] and current sequence alignment algorithms for IDRs cannot easily pick the importance of spacers as well. Although a recent work using machine learning approaches to collect the "features" of IDRs might be an alternative approach to find their traits,[49] it is still limited in capturing the distal interaction, such as the prevailing aromatic residues.

In summary, IDRs appeared late in proteins, and RBPs evolved under selective pressure to assemble precisely in specific cellular locations. This spatiotemporal control can be achieved by mimicking prion properties[50] or by having short α-helical motifs,[32] blocked charged-pattern,[51] or coiled-coil domains,[52] features that provide multivalency and promote the formation of higher-order oligomers. Aromatic residues, with weak π-π or cation-π interactions, appear to have been selected in this role, affording the weak, reversible interactions required for biomolecular condensate formation. After contributing a first time in the evolution of folded proteins, the subsequent return of the (aromatic) rings in RBPs appears to have been crucial to the emergence of LLPS.

## ACKNOWLEDGMENTS

This research was funded by the Ministry of Science and Technology of Taiwan, grant numbers 109-2113-M-010-003 and 110-2113-M-A49A-504-MY3.

## AUTHOR CONTRIBUTIONS

**Wen-Lin Ho:** Data curation (lead); formal analysis (lead); investigation (equal); methodology (equal); software (lead); visualization (equal); writing – review and editing (supporting). **Jie-rong Huang:** Conceptualization (lead); formal analysis (equal); funding acquisition (lead); investigation (equal); methodology (equal); project administration (lead); resources (lead); software (supporting); supervision (lead); validation (equal); visualization (equal); writing – original draft (lead); writing – review and editing (lead).

## DATA AVAILABILITY STATEMENT

All scripts and data generated or analyzed is this study are available in the repository: http://github.com/allmwh/the_return_of_the_rings.

## ORCID

*Jie-rong Huang* https://orcid.org/0000-0003-3674-2228

## REFERENCES

1. Burley SK, Petsko GA. Aromatic-aromatic interaction: A mechanism of protein structure stabilization. Science. 1985;229:23–28.
2. Dunker AK, Lawson JD, Brown CJ, et al. Intrinsically disordered protein. J Mol Graph Model. 2001;19:26–59.
3. Yan J, Cheng J, Kurgan L, Uversky VN. Structural and functional analysis of "non-smelly" proteins. Cell Mol Life Sci. 2020;77:2423–2440.
4. Li HR, Chiang WC, Chou PC, Wang WJ, Huang JR. TAR DNA-binding protein 43 (TDP-43) liquid-liquid phase separation is mediated by just a few aromatic residues. J Biol Chem. 2018;293:6090–6098.
5. Lin Y, Currie SL, Rosen MK. Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs. J Biol Chem. 2017;292:19110–19120.
6. Kwon I, Kato M, Xiang S, et al. Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains. Cell. 2013;155:1049–1060.
7. Wang J, Choi JM, Holehouse AS, et al. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. Cell. 2018;174:688–699.e16.
8. Eck RV, Dayhoff MO. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. Science. 1966;152:363–366.
9. Jeffery CJ. Moonlighting proteins. Trends Biochem Sci. 1999;24:8–11.
10. Oldfield CJ, Dunker AK. Intrinsically disordered proteins and intrinsically disordered protein regions. Annu Rev Biochem. 2014;83:553–584.
11. Smock RG, Gierasch LM. Sending signals dynamically. Science. 2009;324:198–203.
12. Alberti S, Hyman AA. Biomolecular condensates at the nexus of cellular stress, protein aggregation disease and ageing. Nat Rev Mol Cell Biol. 2021;22:196–213.
13. Trifonov EN. The triplet code from first principles. J Biomol Struct Dyn. 2004;22:1–11.

14. Katsnelson A. Did disordered proteins help launch life on earth? ACS Cent Sci. 2020;6:1854–1857.
15. Kulkarni P, Uversky VN. Intrinsically disordered proteins: The dark horse of the dark proteome. Proteomics. 2018;18: e1800061.
16. Zhu H, Sepulveda E, Hartmann MD, et al. Origin of a folded repeat protein from an intrinsically disordered ancestor. Elife. 2016;5:e16761.
17. Uversky VN. A decade and a half of protein intrinsic disorder: Biology still waits for physics. Protein Sci. 2013;22:693–724.
18. Mohan A, Oldfield CJ, Radivojac P, et al. Analysis of molecular recognition features (MoRFs). J Mol Biol. 2006;362:1043–1059.
19. Vacic V, Oldfield CJ, Mohan A, et al. Characterization of molecular recognition features, MoRFs, and their binding partners. J Proteome Res. 2007;6:2351–2366.
20. Wiedner HJ, Giudice J. It's not just a phase: Function and characteristics of RNA-binding proteins in phase separation. Nat Struct Mol Biol. 2021;28:465–473.
21. Kloc M, Zearfoss NR, Etkin LD. Mechanisms of subcellular mRNA localization. Cell. 2002;108:533–544.
22. Martin KC, Ephrussi A. mRNA localization: Gene expression in the spatial dimension. Cell. 2009;136:719–730.
23. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nat Rev Genet. 2014;15:829–845.
24. UniProt C. UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47:D506–D515.
25. Romero O, Dunker K. Sequence data analysis for long disordered regions prediction in the calcineurin family. Genome Inform Ser Workshop Genome Inform. 1997;8:110–124.
26. Radivojac P, Obradovic Z, Brown CJ, Dunker AK. Prediction of boundaries between intrinsically ordered and disordered protein regions. Pac Symp Biocomput. 2003;216–227.
27. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop Genome Inform. 2000;11:161–171.
28. Zagrovic B, Bartonek L, Polyansky AA. RNA-protein interactions in an unstructured context. FEBS Lett. 2018;592:2901–2916.
29. Varadi M, Zsolyomi F, Guharoy M, Tompa P. Functional advantages of conserved intrinsic disorder in RNA-binding proteins. PLoS One. 2015;10:e0139731.
30. Zhao B, Katuwawala A, Oldfield CJ, et al. Intrinsic disorder in human RNA-binding proteins. J Mol Biol. 2021;433:167229.
31. Capra JA, Singh M. Predicting functionally important residues from sequence conservation. Bioinformatics. 2007;23:1875–1882.
32. Conicella AE, Zerze GH, Mittal J, Fawzi NL. ALS mutations disrupt phase separation mediated by alpha-helical structure in the TDP-43 low-complexity C-terminal domain. Structure. 2016;24:1537–1549.
33. McGurk L, Gomes E, Guo L, et al. Poly(ADP-ribose) prevents pathological phase separation of TDP-43 by promoting liquid demixing and stress granule localization. Mol Cell. 2018;71: 703–717.e9.
34. Babinchak WM, Haider R, Dumm BK, et al. The role of liquid-liquid phase separation in aggregation of the TDP-43 low-complexity domain. J Biol Chem. 2019;294:6306–6317.
35. Sun Y, Medina Cruz A, Hadley KC, et al. Physiologically important electrolytes as regulators of TDP-43 aggregation and droplet-phase behavior. Biochemistry. 2019;58:590–607.
36. Li HR, Chen TC, Hsiao CL, Shi L, Chou CY, Huang JR. The physical forces mediating self-association and phase-separation in the C-terminal domain of TDP-43. Biochim Biophys Acta. 2018;1866:214–223.
37. Chen TC, Hsiao CL, Huang SJ, Huang JR. The nearest-neighbor effect on random-coil NMR chemical shifts demonstrated using a low-complexity amino-acid sequence. Protein Pept Lett. 2016;23:967–975.
38. Qamar S, Wang GZ, Randle SJ, et al. FUS phase separation is modulated by a molecular chaperone and methylation of arginine cation-π interactions. Cell. 2018;173:720–734 e15.
39. Dasmeh P, Wagner A. Natural selection on the phase-separation properties of FUS during 160 my of mammalian evolution. Mol Biol Evol. 2021;38:940–951.
40. Milles S, Mercadante D, Aramburu IV, et al. Plasticity of an ultrafast interaction between nucleoporins and nuclear transport receptors. Cell. 2015;163:734–745.
41. Onischenko E, Tang JH, Andersen KR, et al. Natively unfolded FG repeats stabilize the structure of the nuclear pore complex. Cell. 2017;171:904–917.e19.
42. Hayama R, Sparks S, Hecht LM, et al. Thermodynamic characterization of the multivalent interactions underlying rapid and selective translocation through the nuclear pore complex. J Biol Chem. 2018;293:4555–4563.
43. Bastos R, Lin A, Enarson M, Burke B. Targeting and function in mRNA export of nuclear pore complex protein Nup153. J Cell Biol. 1996;134:1141–1156.
44. Sheu-Gruttadauria J, MacRae IJ. Phase transitions in the assembly and function of human miRISC. Cell. 2018;173:946–957.e16.
45. Ford L, Ling E, Kandel ER, Fioriti L. CPEB3 inhibits translation of mRNA targets by localizing them to P bodies. Proc Natl Acad Sci USA. 2019;116:18078–18087.
46. Batlle C, Yang P, Coughlin M, et al. hnRNPDL phase separation is regulated by alternative splicing and disease-causing mutations accelerate its aggregation. Cell Rep. 2020;30:1117–1128.e5.
47. Martin EW, Holehouse AS, Peran I, et al. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. Science. 2020;367:694–699.
48. Dasmeh P, Doronin R, Wagner A. The length scale of multivalent interactions is evolutionarily conserved in fungal and vertebrate phase-separating proteins. Genetics. 2022.220, iyab184. https://academic.oup.com/genetics/article/220/1/iyab184/6410649
49. Lu AX, Lu AX, Pritišanac P, Zarin T, Forman-Kay JD, Moses AM. Discovering molecular features of intrinsically disordered regions by using evolution for contrastive learning. bioRxiv. 2021. https://doi.org/10.1101/2021.07.29.454330
50. Patel A, Lee HO, Jawerth L, et al. A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. Cell. 2015;162:1066–1077.
51. Greig JA, Nguyen TA, Lee M, et al. Arginine-enriched mixed-charge domains provide cohesion for nuclear speckle condensation. Mol Cell. 2020;77:1237–1250.e4.

52. Fang X, Wang L, Ishikawa R, et al. Arabidopsis FLL2 promotes liquid-liquid phase separation of polyadenylation complexes. Nature. 2019;569:265–269.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Ho W-L, Huang J. The return of the rings: Evolutionary convergence of aromatic residues in the intrinsically disordered regions of RNA-binding proteins for liquid–liquid phase separation. Protein Science. 2022;31(5): e4317. https://doi.org/10.1002/pro.4317

Supplementary Materials

# The return of the rings: evolutionary convergence of aromatic residues in the intrinsically disordered regions of RNA-binding proteins for liquid-liquid phase separation

Wen-Lin Ho and Jie-rong Huang

## Methods

*Disorder prevalence and disorder odds ratios relative to the human proteome*

All human protein sequences were retrieved from UniProt (UniProtKB_2021_01, download date: 2021/02/06, 20396 sequences in total). Disorder predictions were performed using the VLXT, VL3, and VSL2 algorithms on the PONDR webserver[1,2] and flDPnn.[3] According to these algorithms, each residue in a sequence was annotated as ordered or disordered. The criteria of consecutive disordered regions (longer than 30, 40, or 50 residues) or the type of amino acids within these regions were analyzed using in-house scripts (deposited in GitHub). RNA-binding proteins were based on the definition of the consensus studies.[4] Sub-groups of RNA-binding proteins, targeting to mRNA, irRNA, tRNA, rRNA, ncRNA, snRNA, are defined based on the same study (the nomenclature follows the recent analysis[5]). Names of these RBPs and sequences used in this analysis are in an Excel file in the extended datasets. Nine gene names in the census study are not inconsistent with the UniProt and thus were not analyzed (also indicated in the Excel file). Missing these sequences does not affect our analysis.

*Sequence conservation analysis*

Human RBP orthologues were obtained from the Orthologous Matrix (OMA) database [6] using the human protein's UniProt ID as input to grep its OMA group members. The orthologues were filtered by taxonomy. The sequences of the orthologues were aligned using the Clustal Omega (v.1.2.4)[7] module in Biopython[8]. The aligned sequence with a gap longer than 20 consecutive residues will be removed unless more than three orthologs in the alignment have the same gap. The aligned sequences were mapped to the human sequence's residue number, and the level of conservation of each residue was quantified using the Jensen-Shannon divergence score ($D^{JS}$)[9]:

$$D^{JS} = \frac{1}{2}\sum_{\alpha \in \text{a.a}} p_C(\alpha)\log_2 \frac{p_C(\alpha)}{r(\alpha)} + \frac{1}{2}\sum_{\alpha \in \text{a.a}} q(\alpha)\log_2 \frac{q(\alpha)}{r(\alpha)}$$

where α is one of the twenty amino acids (a.a.), $p_C(\alpha)$ is the column population of amino-acid α, $q(\alpha)$ is the background distribution in the BLOSUM62 alignment, $r(\alpha) = 0.5\ p_C(\alpha) + 0.5\ q(\alpha)$. The orthologs used for the analysis in Figure 2 and Figure S5 are listed in Table S8 or deposited the Github server. The Jensen-Shannon divergence scores were converted using the Z-score function. The mean and standard deviation derived from the chordates were used for normalization.

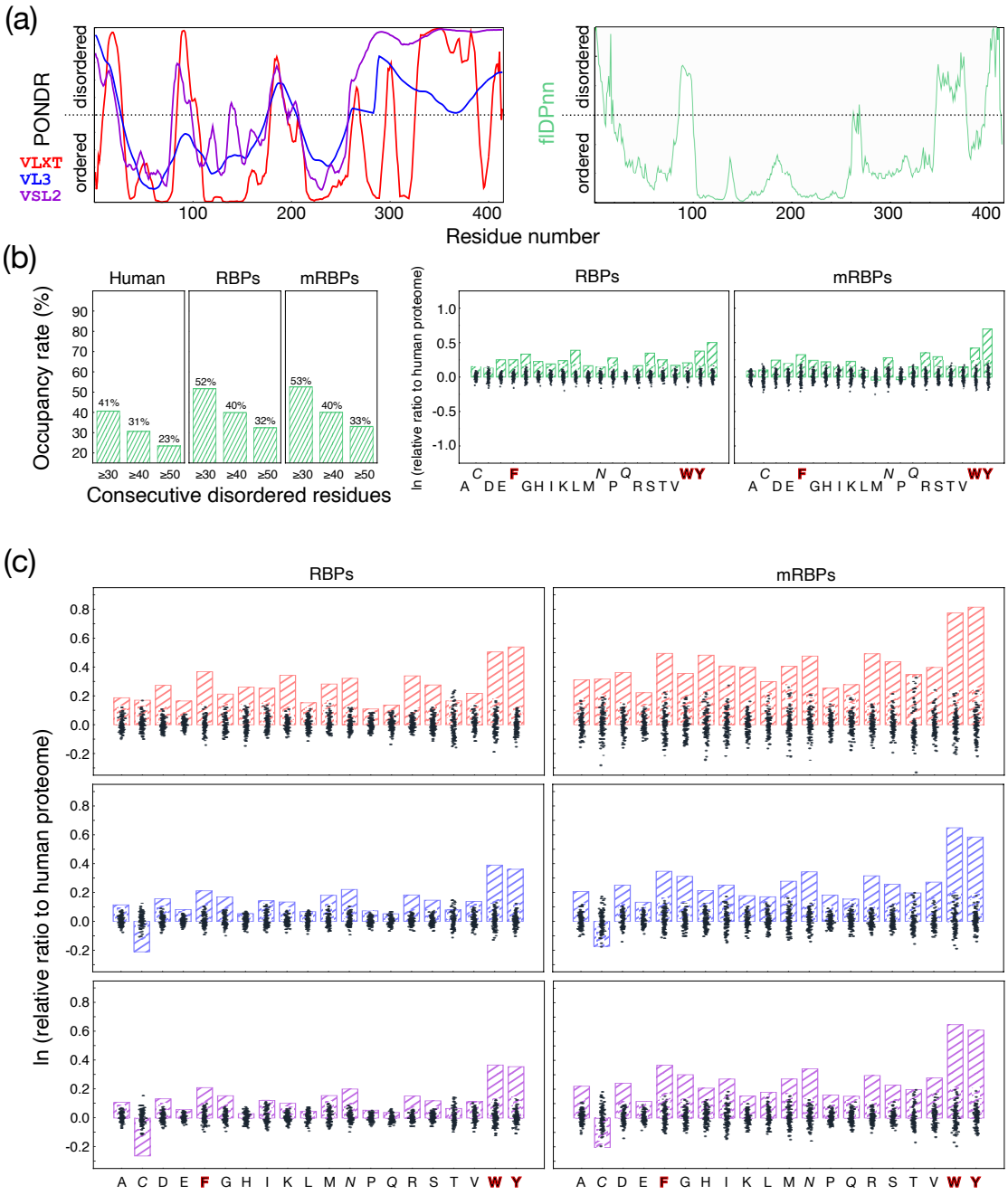


**Figure S1**. The analysis is independent of the algorithms chosen or the length criteria. One of the latest IDP predictors, flDPnn[3], which shows good performance[10] is included for comparison. (a) Examples of using different algorithms: three from the PONDR server (left) and flDPnn (right) on a single protein (TDP-43). (b) Compared to Figure 1, using other algorithms such as the latest flDPnn or (c) analyzed with criteria of consecutive lengths (in this example, ≥ 40 residues). The overall trends are similar.

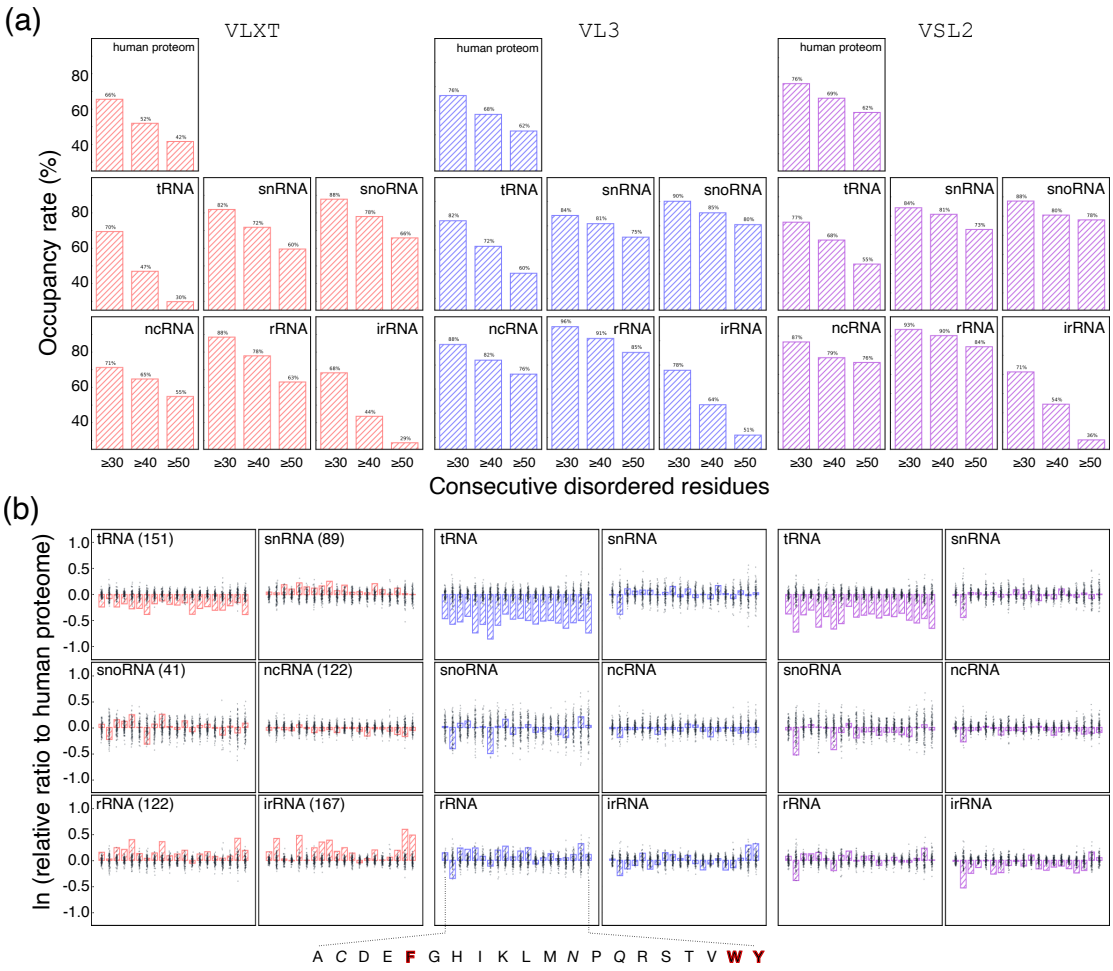


**Figure S2.** Prevalence of disorder and disorder odds ratios relative to the human proteome by amino acid type for different RNA-binding proteins. (a) The proportion of proteins with disordered regions longer than 30, 40, or 50 consecutive residues, as predicted using different algorithms, in the human proteome (upper panel) and other six groups of protein targeting to different types of RNAs based on the census study.[5] (b) Log-odds ratios relative to the human proteome of being in an IDR for amino acids in different types of RNA targeting proteins. The dots represent the values obtained for randomly selected (negative control) subsamples of the human proteome. The number of random sample sizes depends on the RBPs' number in the census data,[4] as indicated.

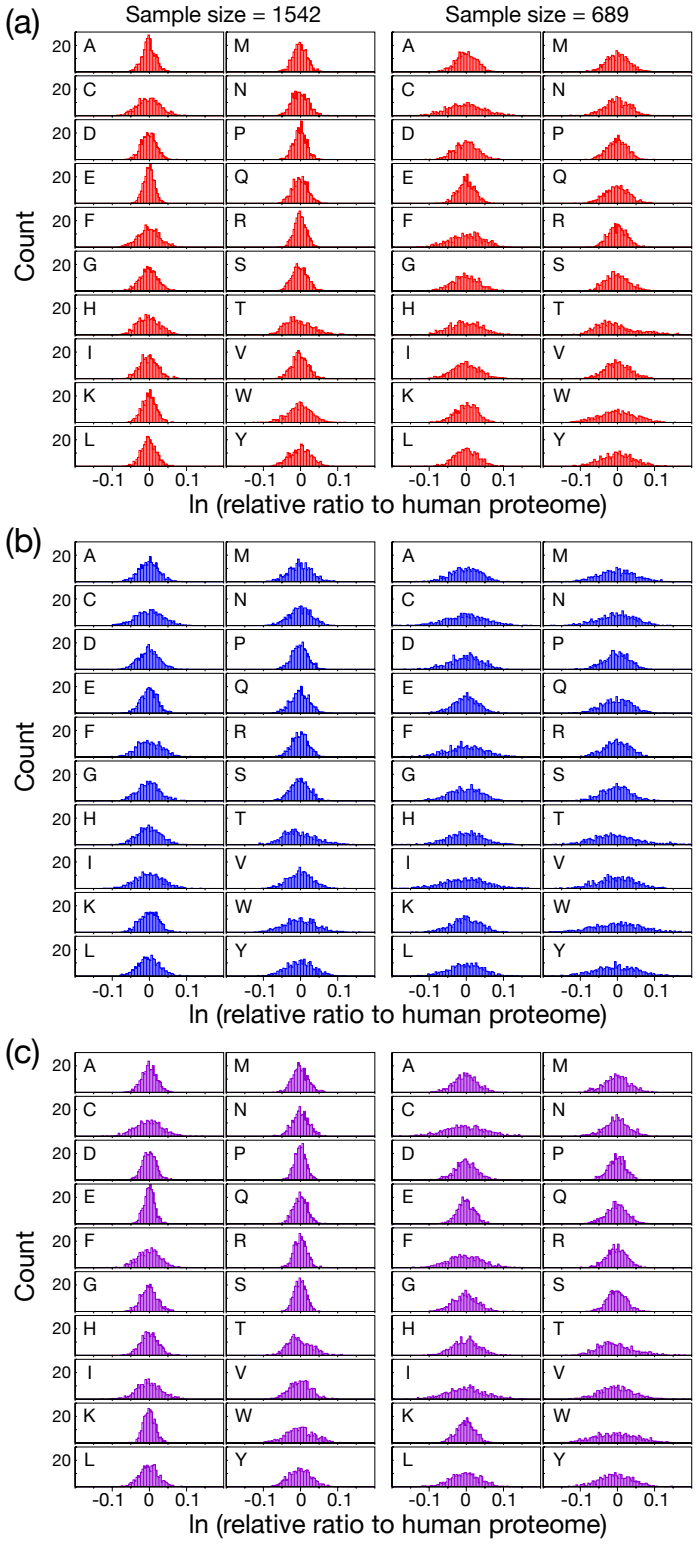


**Figure S3.** Distribution of the log-odds ratio relative to the human proteome of being in an intrinsically disordered region as predicted by the (a) VLXT, (b) VL3, and (c) VSL2 algorithm for randomly selected subgroups of the human proteome (left, N =1542; right, N = 689)

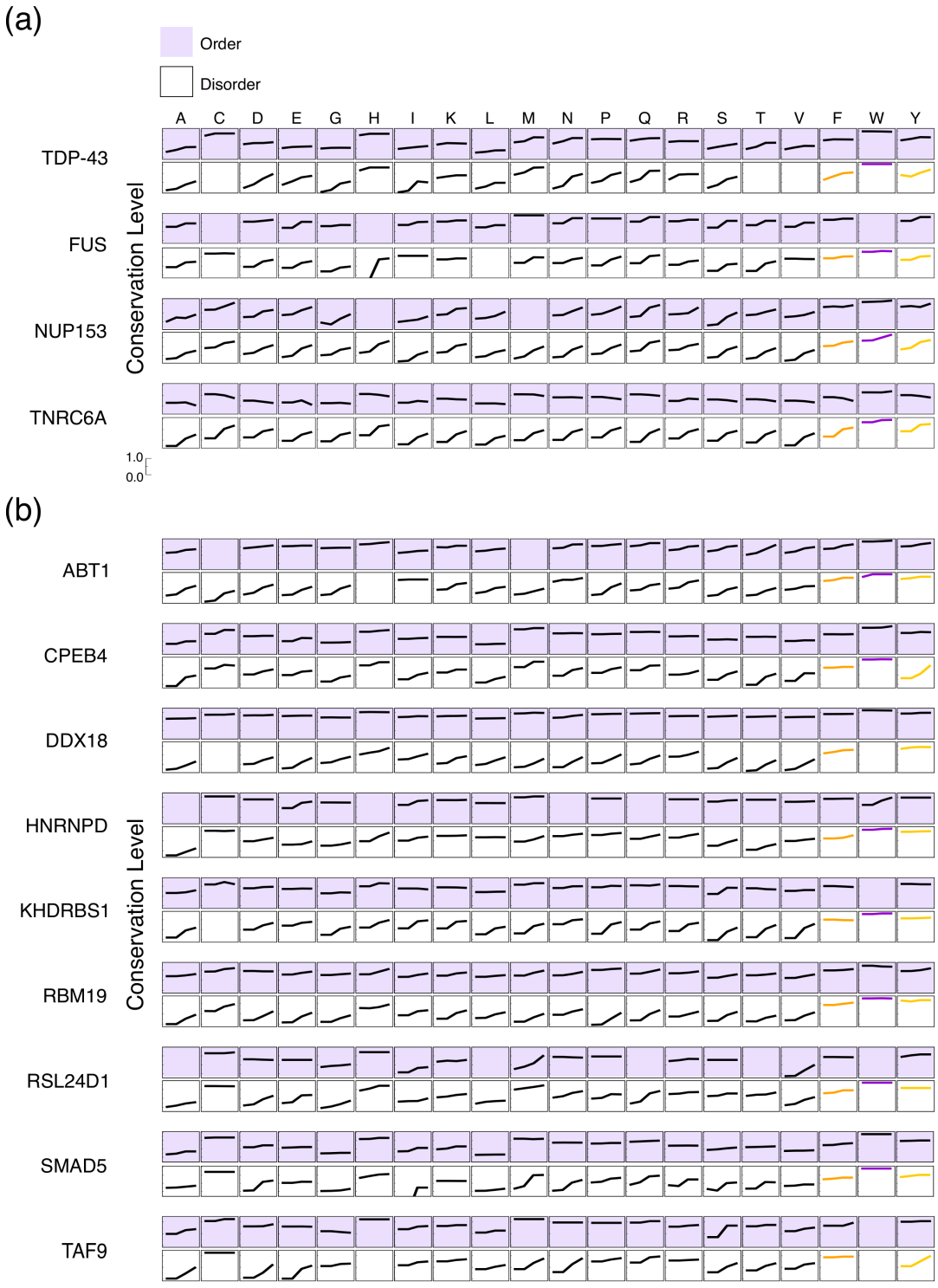


**Figure S4.** Average level of sequence conservation by amino acid type for residues in ordered (purple) or disordered regions (white) as a function of decreasing taxonomic ranks (Chordata, Vertebrata, Tetrapoda, Mammalia), for (a) the example proteins presented in fig. 2 of the main text and (b) the example proteins presented in figure S3 (below). Only residues in disordered (ordered) regions longer than 40 (10) consecutive amino acids were considered.

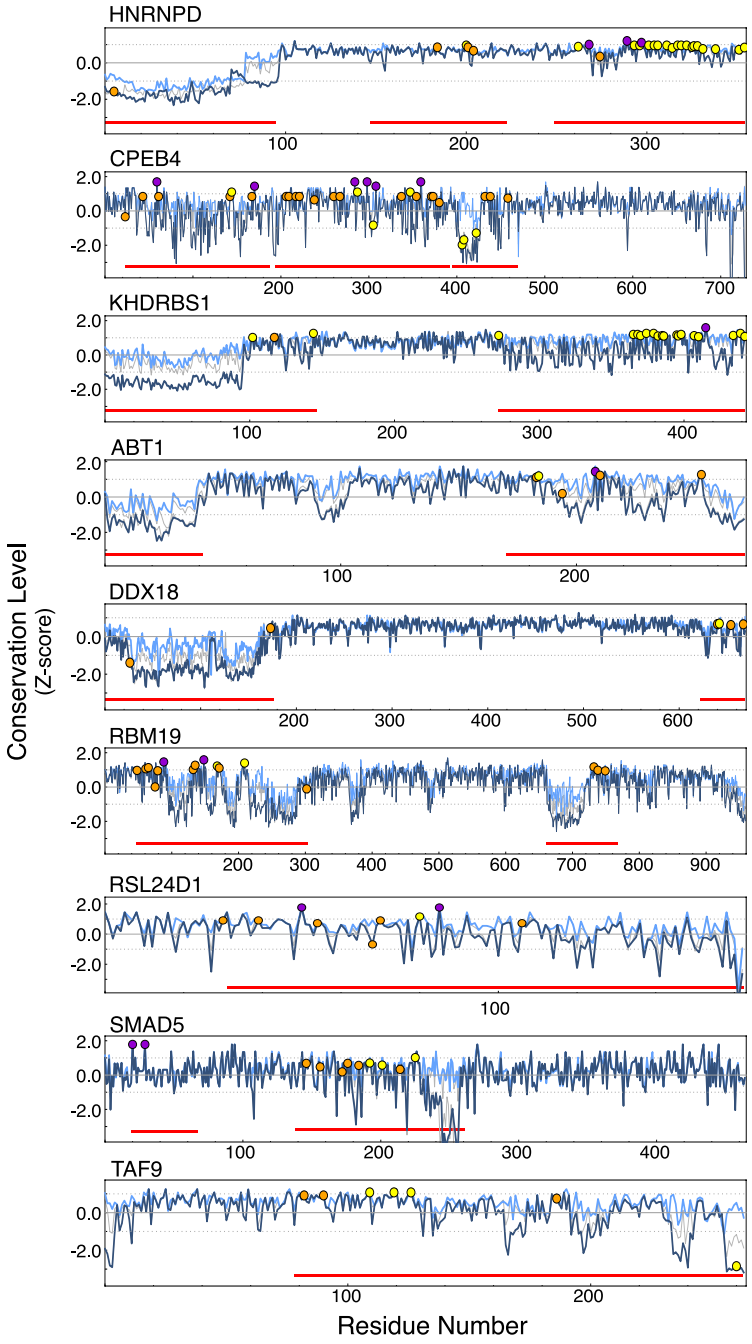


**Figure S5.** Sequence conservation in example proteins with highly conserved aromatic residues. Levels of sequence conservation are quantified by the Jensen-Shannon divergence score and normalized using the Z-score function (the mean of all values toward chordates is 0; the value in the y-axes is the standard deviation; positive values mean more conserved.) Levels of conservation in chordates (dark blue), vertebrates (gray), tetrapods (gray), and mammals (light blue) are plotted versus the corresponding residue number in the human sequence. Predicted disordered regions longer than 40 residues are indicated with red bars. Aromatic residues are labeled on the chordate line: Phe (orange), Trp (purple), Tyr (yellow).

**Table S1.** Proportions of proteins with predicted IDRs longer than certain lengths.

| Algorithm | Length longer than | Human proteins (%) | RNA binding proteins (%) | (RBP – Human) (%) | mRNA binding proteins (%) | (mRBP – Human) (%) |
|---|---|---|---|---|---|---|
| VLXT | **20** | **81.8** | **92.6** | 10.9 | **93.6** | 11.8 |
| | **30** | **65.6** | **79.5** | 13.9 | **84.5** | 18.9 |
| | **40** | **52.0** | **66.2** | 14.2 | **73.0** | 21.0 |
| | **50** | **41.7** | **54.7** | 12.9 | **62.6** | 20.8 |
| | **60** | **33.7** | **45.9** | 12.2 | **55.4** | 21.7 |
| VL3 | **20** | **83.3** | **93.4** | 10.1 | **93.9** | 10.6 |
| | **30** | **76.3** | **88.0** | 11.7 | **90.1** | 13.8 |
| | **40** | **68.5** | **80.9** | 12.4 | **84.3** | 15.8 |
| | **50** | **61.6** | **74.4** | 12.8 | **79.7** | 18.1 |
| | **60** | **55.7** | **66.6** | 10.9 | **74.2** | 18.5 |
| VSL2 | **20** | **84.1** | **91.8** | 7.7 | **93.9** | 9.8 |
| | **30** | **76.5** | **85.3** | 8.8 | **88.4** | 11.9 |
| | **40** | **68.9** | **78.1** | 9.2 | **82.6** | 13.7 |
| | **50** | **61.5** | **71.2** | 9.7 | **78.7** | 17.1 |
| | **60** | **55.6** | **64.3** | 8.7 | **73.4** | 17.7 |

**Table S2.** Number of residues predicted to be in ordered/disordered regions by the VLXT algorithm

| | RBPs | | mRBPs | | Human proteins | |
|---|---|---|---|---|---|---|
| | Ordered | Disordered | Ordered | Disordered | Ordered | Disordered |
| A | 33767 | 27760 | 15945 | 14705 | 474415 | 321897 |
| C | 12353 | 2662 | 5358 | 1244 | 220009 | 41333 |
| D | 27224 | 18481 | 13262 | 9798 | 349268 | 188768 |
| E | 31856 | 36215 | 15343 | 18489 | 409580 | 396968 |
| F | 25340 | 4883 | 12221 | 2586 | 362796 | 51625 |
| G | 32066 | 26379 | 15778 | 15089 | 452912 | 293467 |
| H | 15438 | 6544 | 6995 | 3577 | 220858 | 76954 |
| I | 26308 | 11070 | 12352 | 5838 | 369522 | 123116 |
| K | 34806 | 26359 | 16136 | 12385 | 417435 | 233546 |
| L | 54425 | 25927 | 24430 | 12995 | 802230 | 329184 |
| M | 10887 | 8169 | 5332 | 4419 | 151790 | 90374 |
| N | 21112 | 11254 | 10520 | 6336 | 287931 | 119746 |
| P | 25046 | 31437 | 11796 | 18099 | 341273 | 375933 |
| Q | 25311 | 18377 | 12474 | 10261 | 330112 | 211467 |
| R | 26230 | 30610 | 12211 | 16772 | 347223 | 292901 |
| S | 34805 | 36396 | 16745 | 21500 | 529363 | 416555 |
| T | 26013 | 17415 | 12199 | 9710 | 388457 | 219237 |
| V | 34359 | 17768 | 15600 | 9331 | 474983 | 202342 |
| W | 6853 | 2006 | 2992 | 1062 | 114749 | 23332 |
| Y | 18116 | 5406 | 8934 | 3333 | 251740 | 50805 |

**Table S3.** Number of residues predicted to be in ordered/disordered regions by the VL3 algorithm

| | RBPs | | mRBPs | | Human proteins | |
|---|---|---|---|---|---|---|
| | Ordered | Disordered | Ordered | Disordered | Ordered | Disordered |
| A | 34295 | 27232 | 16011 | 14639 | 479203 | 317109 |
| C | 10891 | 4124 | 4765 | 1837 | 175643 | 85699 |
| D | 25476 | 20229 | 11988 | 11072 | 330080 | 207956 |
| E | 31199 | 36872 | 14727 | 19105 | 399940 | 406608 |
| F | 22220 | 8003 | 10423 | 4384 | 324013 | 90408 |
| G | 29712 | 28733 | 13694 | 17173 | 431350 | 315029 |
| H | 13748 | 8234 | 6048 | 4524 | 191475 | 106337 |
| I | 27452 | 9926 | 12931 | 5259 | 378838 | 113800 |
| K | 30335 | 30830 | 13731 | 14790 | 360043 | 290938 |
| L | 53010 | 27342 | 23653 | 13772 | 771505 | 359909 |
| M | 11504 | 7552 | 5595 | 4156 | 159982 | 82182 |
| N | 19550 | 12816 | 9461 | 7395 | 274701 | 132976 |
| P | 23244 | 33239 | 10572 | 19323 | 320476 | 396730 |
| Q | 22870 | 20818 | 10930 | 11805 | 295531 | 246048 |
| R | 27151 | 29689 | 12085 | 16898 | 355342 | 284782 |
| S | 31099 | 40102 | 14545 | 23700 | 477217 | 468701 |
| T | 25201 | 18227 | 11733 | 10176 | 370569 | 237125 |
| V | 34809 | 17318 | 15676 | 9255 | 478562 | 198763 |
| W | 6619 | 2240 | 2768 | 1286 | 113612 | 24469 |
| Y | 16578 | 6944 | 7858 | 4409 | 238289 | 64256 |

**Table S4.** Number of residues predicted to be in ordered/disordered regions by the VLS2 algorithm

| | RBPs | | mRBPs | | Human proteins | |
|---|---|---|---|---|---|---|
| | Ordered | Disordered | Ordered | Disordered | Ordered | Disordered |
| A | 30642 | 30885 | 13975 | 16675 | 429548 | 366764 |
| C | 11246 | 3769 | 4884 | 1718 | 178491 | 82851 |
| D | 22064 | 23641 | 10146 | 12914 | 281602 | 256434 |
| E | 26089 | 41982 | 12175 | 21657 | 326165 | 480383 |
| F | 21578 | 8645 | 10046 | 4761 | 313891 | 100530 |
| G | 25897 | 32548 | 11723 | 19144 | 374088 | 372291 |
| H | 12557 | 9425 | 5427 | 5145 | 170785 | 127027 |
| I | 27085 | 10293 | 12597 | 5593 | 370211 | 122427 |
| K | 25112 | 36053 | 11275 | 17246 | 290910 | 360071 |
| L | 52104 | 28248 | 22962 | 14463 | 750188 | 381226 |
| M | 10076 | 8980 | 4834 | 4917 | 138373 | 103791 |
| N | 16953 | 15413 | 7988 | 8868 | 237124 | 170553 |
| P | 19526 | 36957 | 8703 | 21192 | 263731 | 453475 |
| Q | 19952 | 23736 | 9358 | 13377 | 255138 | 286441 |
| R | 24648 | 32192 | 10737 | 18246 | 314028 | 326096 |
| S | 24465 | 46736 | 11073 | 27172 | 373370 | 572548 |
| T | 22428 | 21000 | 10245 | 11664 | 328426 | 279268 |
| V | 34945 | 17182 | 15635 | 9296 | 473431 | 203894 |
| W | 6355 | 2504 | 2622 | 1432 | 108987 | 29094 |
| Y | 15972 | 7550 | 7451 | 4816 | 228179 | 74366 |

**Table S5**. Summary statistics for the log-odds ratios relative to the human proteome of being in a disordered region as predicted by the VXLT algorithm for randomly sampled subgroups of the human proteome

| | Random sample size = 1542 | | | | Random sample size =689 | | | |
|---|---|---|---|---|---|---|---|---|
| | mean | std | max | min | mean | std | max | min |
| A | 0.00 | 0.02 | 0.05 | -0.05 | 0.00 | 0.03 | 0.08 | -0.09 |
| C | 0.00 | 0.03 | 0.10 | -0.09 | 0.00 | 0.05 | 0.15 | -0.15 |
| D | 0.00 | 0.02 | 0.08 | -0.06 | 0.00 | 0.03 | 0.09 | -0.11 |
| E | 0.00 | 0.01 | 0.04 | -0.05 | 0.00 | 0.02 | 0.07 | -0.08 |
| F | 0.00 | 0.03 | 0.08 | -0.08 | 0.00 | 0.04 | 0.12 | -0.13 |
| G | 0.00 | 0.02 | 0.08 | -0.07 | 0.00 | 0.04 | 0.13 | -0.12 |
| H | 0.00 | 0.03 | 0.08 | -0.08 | 0.00 | 0.04 | 0.18 | -0.13 |
| I | 0.00 | 0.02 | 0.08 | -0.06 | 0.00 | 0.04 | 0.13 | -0.10 |
| K | 0.00 | 0.02 | 0.06 | -0.06 | 0.00 | 0.03 | 0.09 | -0.10 |
| L | 0.00 | 0.02 | 0.06 | -0.07 | 0.00 | 0.03 | 0.10 | -0.11 |
| M | 0.00 | 0.02 | 0.08 | -0.06 | 0.00 | 0.03 | 0.11 | -0.08 |
| N | 0.00 | 0.02 | 0.07 | -0.07 | 0.00 | 0.03 | 0.11 | -0.13 |
| P | 0.00 | 0.02 | 0.06 | -0.05 | 0.00 | 0.03 | 0.10 | -0.09 |
| Q | 0.00 | 0.02 | 0.07 | -0.06 | 0.00 | 0.03 | 0.09 | -0.12 |
| R | 0.00 | 0.02 | 0.05 | -0.06 | 0.00 | 0.02 | 0.09 | -0.08 |
| S | 0.00 | 0.02 | 0.07 | -0.06 | 0.00 | 0.03 | 0.11 | -0.11 |
| T | 0.00 | 0.04 | 0.16 | -0.09 | 0.00 | 0.06 | 0.22 | -0.14 |
| V | 0.00 | 0.02 | 0.06 | -0.07 | 0.00 | 0.03 | 0.09 | -0.11 |
| W | 0.00 | 0.03 | 0.12 | -0.13 | 0.00 | 0.05 | 0.15 | -0.13 |
| Y | 0.00 | 0.03 | 0.09 | -0.08 | 0.00 | 0.04 | 0.13 | -0.12 |

The values are derived from 1000 randomly selected subgroups of the human proteome with sample sizes of 1542 or 689 (the same as those of the samples of RBPs and mRBPs considered) using eq. (1). The mean, standard deviation (std), maximum (max), and minimum (min) values are listed.

**Table S6**. Summary statistics for the log-odds ratios relative to the human proteome of being in a disordered region as predicted by the VL3 algorithm for randomly sampled subgroups of the human proteome

| | Random sample size = 1542 | | | | Random sample size =689 | | | |
|---|---|---|---|---|---|---|---|---|
| | mean | std | max | min | mean | std | max | min |
| A | 0.00 | 0.03 | 0.08 | -0.08 | 0.00 | 0.04 | 0.11 | -0.13 |
| C | 0.00 | 0.04 | 0.11 | -0.10 | 0.00 | 0.06 | 0.17 | -0.19 |
| D | 0.00 | 0.03 | 0.10 | -0.08 | 0.00 | 0.04 | 0.11 | -0.17 |
| E | 0.00 | 0.02 | 0.06 | -0.07 | 0.00 | 0.03 | 0.10 | -0.11 |
| F | 0.00 | 0.04 | 0.14 | -0.12 | 0.00 | 0.05 | 0.14 | -0.17 |
| G | 0.00 | 0.03 | 0.12 | -0.11 | 0.00 | 0.04 | 0.10 | -0.16 |
| H | 0.00 | 0.03 | 0.10 | -0.10 | 0.00 | 0.05 | 0.13 | -0.19 |
| I | 0.00 | 0.04 | 0.14 | -0.12 | 0.00 | 0.06 | 0.17 | -0.19 |
| K | 0.00 | 0.03 | 0.08 | -0.07 | 0.00 | 0.04 | 0.11 | -0.14 |
| L | 0.00 | 0.03 | 0.10 | -0.08 | 0.00 | 0.05 | 0.14 | -0.14 |
| M | 0.00 | 0.03 | 0.09 | -0.09 | 0.00 | 0.05 | 0.14 | -0.17 |
| N | 0.00 | 0.03 | 0.09 | -0.09 | 0.00 | 0.05 | 0.14 | -0.18 |
| P | 0.00 | 0.02 | 0.06 | -0.06 | 0.00 | 0.03 | 0.09 | -0.11 |
| Q | 0.00 | 0.02 | 0.08 | -0.07 | 0.00 | 0.04 | 0.13 | -0.13 |
| R | 0.00 | 0.02 | 0.07 | -0.09 | 0.00 | 0.04 | 0.10 | -0.13 |
| S | 0.00 | 0.03 | 0.08 | -0.07 | 0.00 | 0.04 | 0.12 | -0.13 |
| T | 0.00 | 0.04 | 0.14 | -0.11 | -0.01 | 0.06 | 0.21 | -0.19 |
| V | 0.00 | 0.03 | 0.12 | -0.10 | 0.00 | 0.05 | 0.13 | -0.17 |
| W | 0.00 | 0.04 | 0.16 | -0.12 | 0.00 | 0.07 | 0.20 | -0.23 |
| Y | 0.00 | 0.04 | 0.12 | -0.12 | 0.00 | 0.06 | 0.16 | -0.21 |

The values are derived from 1000 randomly selected subgroups of the human proteome with sample sizes of 1542 or 689 (the same as those of the samples of RBPs and mRBPs considered) using eq. (1). The mean, standard deviation (std), maximum (max), and minimum (min) values are listed.

**Table S7**. Summary statistics for the log-odds ratios relative to the human proteome of being in a disordered region as predicted by the VSL2 algorithm for randomly sampled subgroups of the human proteome

| | Random sample size = 1542 | | | | Random sample size =689 | | | |
|---|---|---|---|---|---|---|---|---|
| | mean | std | max | min | mean | std | max | min |
| A | 0.00 | 0.02 | 0.06 | -0.07 | 0.00 | 0.03 | 0.11 | -0.09 |
| C | 0.00 | 0.04 | 0.14 | -0.13 | 0.00 | 0.06 | 0.14 | -0.21 |
| D | 0.00 | 0.02 | 0.07 | -0.05 | 0.00 | 0.03 | 0.12 | -0.10 |
| E | 0.00 | 0.01 | 0.05 | -0.05 | 0.00 | 0.02 | 0.07 | -0.08 |
| F | 0.00 | 0.03 | 0.10 | -0.11 | 0.00 | 0.05 | 0.15 | -0.14 |
| G | 0.00 | 0.02 | 0.07 | -0.07 | 0.00 | 0.03 | 0.12 | -0.10 |
| H | 0.00 | 0.02 | 0.08 | -0.07 | 0.00 | 0.04 | 0.12 | -0.11 |
| I | 0.00 | 0.03 | 0.10 | -0.13 | 0.00 | 0.05 | 0.16 | -0.16 |
| K | 0.00 | 0.02 | 0.05 | -0.05 | 0.00 | 0.03 | 0.10 | -0.07 |
| L | 0.00 | 0.03 | 0.09 | -0.08 | 0.00 | 0.04 | 0.15 | -0.12 |
| M | 0.00 | 0.02 | 0.07 | -0.07 | 0.00 | 0.03 | 0.11 | -0.10 |
| N | 0.00 | 0.02 | 0.06 | -0.08 | 0.00 | 0.03 | 0.10 | -0.11 |
| P | 0.00 | 0.02 | 0.05 | -0.05 | 0.00 | 0.02 | 0.09 | -0.08 |
| Q | 0.00 | 0.02 | 0.07 | -0.07 | 0.00 | 0.03 | 0.10 | -0.09 |
| R | 0.00 | 0.02 | 0.07 | -0.05 | 0.00 | 0.03 | 0.08 | -0.10 |
| S | 0.00 | 0.02 | 0.05 | -0.05 | 0.00 | 0.03 | 0.09 | -0.08 |
| T | 0.00 | 0.03 | 0.11 | -0.09 | 0.00 | 0.05 | 0.18 | -0.14 |
| V | 0.00 | 0.03 | 0.10 | -0.10 | 0.00 | 0.05 | 0.18 | -0.13 |
| W | 0.00 | 0.04 | 0.12 | -0.12 | 0.00 | 0.06 | 0.17 | -0.16 |
| Y | 0.00 | 0.03 | 0.12 | -0.10 | 0.00 | 0.05 | 0.15 | -0.16 |

The values are derived from 1000 randomly selected subgroups of the human proteome with sample sizes of 1542 or 689 (the same as those of the samples of RBPs and mRBPs considered) using eq. (1). The mean, standard deviation (std), maximum (max), and minimum (min) values are listed.

**Table 8.** Orthologs used in the analysis of Figure 2.

| OMA Protein ID | Species | OMA cross reference | Index[1] |
|---|---|---|---|
| **TDP-43** | | | |
| BRAFL24270 | *Branchiostoma floridae* | C3ZIE9 | 1 |
| CIOSA12468 | *Ciona savignyi* | H2ZBU8 | 1 |
| LATCH14939 | *Latimeria chalumnae* | H3BC22 | 2 |
| LEPOC09981 | *Lepisosteus oculatus* | W5MHW9 | 2 |
| ANGAN38603 | *Anguilla anguilla* | XP_035237286 | 2 |
| DANRE36783 | *Danio rerio* | A0A2R8Q8T5 | 2 |
| ESOLU39540 | *Esox lucius* | ENSELUG00000009090.1 | 2 |
| SALSA63007 | *Salmo salar* | C0H962 | 2 |
| ORYLA13842 | *Oryzias latipes* | ENSORLG00000003593 | 2 |
| ORENI03731 | *Oreochromis niloticus* | ENSONIG00000004027 | 2 |
| TAKRU15372 | *Takifugu rubripes* | ENSTRUG00000012063 | 2 |
| NEOBR03357 | *Neolamprologus brichardi* | A0A3Q4HV10 | 2 |
| AMPCI13505 | *Amphilophus citrinellus* | A0A3Q0RHE0 | 2 |
| MOLML02764 | *Mola mola* | A0A3Q4B8R2 | 2 |
| NOTFU16587 | *Nothobranchius furzeri* | XP_015814443 | 2 |
| HIPCM06019 | *Hippocampus comes* | A0A3Q2YIA0 | 2 |
| AMPPE26829 | *Amphiprion percula* | A0A3P8RUU9 | 2 |
| XENLA32018 | *Xenopus laevis* | A0A1L8FFB1 | 3 |
| XENTR12916 | *Xenopus tropicalis* | TADBP_XENTR | 3 |
| CHRPI32751 | *Chrysemys picta bellii* | ENSCPBG00000022012.1 | 3 |
| SPHPU16313 | *Sphenodon punctatus* | ENSSPUG00000006688.1 | 3 |
| ANAPL05491 | *Anas platyrhynchos* | ENSAPLG00000006001 | 3 |
| ANAPP02225 | *Anas platyrhynchos platyrhynchos* | U3IED4 | 3 |
| CHICK10318 | *Gallus gallus* | F1NBY1 | 3 |
| MELGA07585 | *Meleagris gallopavo* | G1MZJ1 | 3 |
| SERCA17445 | *Serinus canaria* | ENSSCAG00000016954.1 | 3 |
| PARMJ11107 | *Parus major* | ENSPMJG00000014837.1 | 3 |
| MELUD13787 | *Melopsittacus undulatus* | ENSMUNG00000016204.1 | 3 |
| PELSI14462 | *Pelodiscus sinensis* | K7FJ67 | 3 |
| JUNHY18871 | *Junco hyemalis* | ENSJHYG00000010677.1 | 3 |
| TAEGU06701 | *Taeniopygia guttata* | H0YYI1 | 3 |
| FICAL11802 | *Ficedula albicollis* | U3JXL3 | 3 |
| CHEAB01735 | *Chelonoidis abingdonii* | ENSCABG00000017429.1 | 3 |
| ORNAN13991 | *Ornithorhynchus anatinus* | F7EDX1 | 4 |
| SARHA02491 | *Sarcophilus harrisii* | G3WM39 | 4 |
| DASNO07425 | *Dasypus novemcinctus* | ENSDNOG00000008531 | 4 |
| ERIEU01545 | *Erinaceus europaeus* | ENSEEUG00000003755 | 4 |
| CALJA18973 | *Callithrix jacchus* | F7CL81 | 4 |
| CERAT28967 | *Cercocebus atys* | A0A2K5LQ78 | 4 |
| MACFA23554 | *Macaca fascicularis* | A0A2K5WEW9 | 4 |
| MACMU00269 | *Macaca mulatta* | ENSMMUG00000007456 | 4 |
| MACNE36802 | *Macaca nemestrina* | A0A2K6BKI8 | 4 |
| PAPAN00174 | *Papio anubis* | A0A096NBQ0 | 4 |
| MANLE13078 | *Mandrillus leucophaeus* | A0A2K6A5S4 | 4 |
| PANPA00278 | *Pan paniscus* | A0A2R9C1F8 | 4 |
| HUMAN47411 | *Homo sapiens* | TADBP_HUMAN | 4 |
| CANLF07184 | *Canis lupus familiaris* | ENSCAFG00000016759 | 4 |
| VULVU24847 | *Vulpes vulpes* | ENSVVUG00000027110.1 | 4 |
| URSAM28564 | *Ursus americanus* | A0A452R9Z0 | 4 |
| AILME09206 | *Ailuropoda melanoleuca* | D2HKA5 | 4 |
| FELCA08644 | *Felis catus* | A0A2I2U8C8 | 4 |
| TURTR12812 | *Tursiops truncatus* | ENSTTRG00000001312 | 4 |
| LOXAF11400 | *Loxodonta africana* | G3TD75 | 4 |
| HORSE08842 | *Equus caballus* | F6WAU6 | 4 |
| PROCA06978 | *Procavia capensis* | ENSPCAG00000010442 | 4 |
| PIGXX29002 | *Sus scrofa* | I3LNA4 | 4 |
| BOVIN08303 | *Bos taurus* | G3MX91 | 4 |
| CAPHI06804 | *Capra hircus* | A0A452G5Y7 | 4 |
| MANJA47321 | *Manis javanica* | XP_036855904 | 4 |
| OCHPR06808 | *Ochotona princeps* | ENSOPRG00000006352 | 4 |
| DIPOR02872 | *Dipodomys ordii* | ENSDORG00000007465 | 4 |
| CRIGR20744 | *Cricetulus griseus* | ENSCGRG00000007923.1 | 4 |
| MOUSE42872 | *Mus musculus* | TADBP_MOUSE | 4 |
| CAVPO10357 | *Cavia porcellus* | A0A286XX33 | 4 |
| HETGA00594 | *Heterocephalus glaber* | G5C522 | 4 |
| VICPA09903 | *Vicugna pacos* | ENSVPAG00000009543 | 4 |
| OTOGA09169 | *Otolemur garnettii* | H0XFZ2 | 4 |
| CHILA18976 | *Chinchilla lanigera* | ENSCLAG00000014547.1 | 4 |
| AOTNA20136 | *Aotus nancymaae* | A0A2K5EQY0 | 4 |
| TUPBE02937 | *Tupaia belangeri* | ENSTBEG00000000263 | 4 |
| CAVAP09087 | *Cavia aperea* | ENSCAPG00000015189.1 | 4 |
| PHACI11189 | *Phascolarctos cinereus* | A0A6P5JEZ4 | 4 |
| SAIBB26155 | *Saimiri boliviensis boliviensis* | A0A2K6TZ91 | 4 |
| ICTTR08691 | *Ictidomys tridecemlineatus* | I3MBK5 | 4 |
| JACJA09000 | *Jaculus jaculus* | ENSJJAG00000014196.1 | 4 |
| MYOLU14708 | *Myotis lucifugus* | G1P208 | 4 |
| CHLSB09400 | *Chlorocebus sabaeus* | A0A0D9S8N2 | 4 |

| | | | |
|---|---|---|---|
| RHIBE13964 | *Rhinopithecus bieti* | A0A2K6LE20 | 4 |
| RHIRO33581 | *Rhinopithecus roxellana* | A0A2K6QV39 | 4 |
| EQUAS11333 | *Equus asinus* | ENSEASG00005019017.1 | 4 |
| COLAP32291 | *Colobus angolensis palliatus* | A0A2K5IA07 | 4 |
| FUKDA15705 | *Fukomys damarensis* | A0A091DKG3 | 4 |
| NANGA15865 | *Nannospalax galili* | ENSNGAG00000023416.1 | 4 |
| CARSF02449 | *Carlito syrichta* | A0A1U7SU45 | 4 |

## FUS

| | | | |
|---|---|---|---|
| LEPOC18149 | *Lepisosteus oculatus* | W5LVA3 | 2 |
| ANGAN53300 | *Anguilla anguilla* | XP_035254071 | 2 |
| ANATE28040 | *Anabas testudineus* | A0A3Q1JBK0 | 2 |
| SERDU24279 | *Seriola dumerili* | A0A3B4UPE4 | 2 |
| SCOMX14580 | *Scophthalmus maximus* | ENSSMAG00000011685.1 | 2 |
| TETNG10220 | *Tetraodon nigroviridis* | H3D156 | 2 |
| NOTFU30467 | *Nothobranchius furzeri* | A0A1A7ZC85 | 2 |
| KRYMA15889 | *Kryptolebias marmoratus* | A0A3Q2ZW58 | 2 |
| ASTCA29489 | *Astatotilapia calliptera* | A0A3P8QDK7 | 2 |
| HAPBU09804 | *Haplochromis burtoni* | A0A3Q2W7N5 | 2 |
| NEOBR22012 | *Neolamprologus brichardi* | A0A3Q4H4B7 | 2 |
| AMPOC31093 | *Amphiprion ocellaris* | A0A3Q1B6C3 | 2 |
| AMPPE11049 | *Amphiprion percula* | A0A3P8SA41 | 2 |
| ESOLU17742 | *Esox lucius* | A0A3P8Z6T5 | 2 |
| SALSA21612 | *Salmo salar* | A0A1S3R6J7 | 2 |
| ASTMX07328 | *Astyanax mexicanus* | W5KQF2 | 2 |
| PYGNA09213 | *Pygocentrus nattereri* | A0A3B4E8M4 | 2 |
| ICTPU24291 | *Ictalurus punctatus* | ENSIPUG00000007342.1 | 2 |
| DANRE29660 | *Danio rerio* | F1R0M4 | 2 |
| LATCH06576 | *Latimeria chalumnae* | H3A671 | 2 |
| ECHTE00861 | *Echinops telfairi* | ENSETEG00000015358 | 2 |
| PONAB04910 | *Pongo abelii* | H2NQS4 | 2 |
| CHRPI29771 | *Chrysemys picta bellii* | ENSCPBG00000008979.1 | 3 |
| CHEAB27054 | *Chelonoidis abingdonii* | ENSCABG00000010374.1 | 3 |
| SPHPU24372 | *Sphenodon punctatus* | ENSSPUG00000002998.1 | 3 |
| ANOCA17128 | *Anolis carolinensis* | ENSACAG00000006776 | 3 |
| XENTR14904 | *Xenopus tropicalis* | Q28EL3 | 3 |
| LOXAF05575 | *Loxodonta africana* | G3SNY8 | 4 |
| RABIT15423 | *Oryctolagus cuniculus* | ENSOCUG00000021526.2 | 4 |
| DIPOR01258 | *Dipodomys ordii* | ENSDORG00000014218 | 4 |
| FUKDA20682 | *Fukomys damarensis* | ENSFDAG00000007552.1 | 4 |
| HETGA21202 | *Heterocephalus glaber* | ENSHGLG00100006254.1 | 4 |
| CHILA19792 | *Chinchilla lanigera* | ENSCLAG00000010457.1 | 4 |
| OCTDE00228 | *Octodon degus* | A0A6P3EP53 | 4 |
| JACJA23117 | *Jaculus jaculus* | ENSJJAG00000011802.1 | 4 |
| CRIGR21309 | *Cricetulus griseus* | ENSCGRG00000012106.1 | 4 |
| MOUSE56790 | *Mus musculus* | FUS_MOUSE | 4 |
| RATNO02010 | *Rattus norvegicus* | Q5PQK2 | 4 |
| NANGA16163 | *Nannospalax galili* | ENSNGAG00000023361.1 | 4 |
| ICTTR12793 | *Ictidomys tridecemlineatus* | ENSSTOG00000005943 | 4 |
| CERAT36652 | *Cercocebus atys* | A0A2K5M862 | 4 |
| CHLSB14622 | *Chlorocebus sabaeus* | A0A0D9R082 | 4 |
| MACFA24759 | *Macaca fascicularis* | A0A2K5UY14 | 4 |
| MACMU14458 | *Macaca mulatta* | ENSMMUG00000019637 | 4 |
| MACNE37188 | *Macaca nemestrina* | A0A2K6B648 | 4 |
| MANLE13941 | *Mandrillus leucophaeus* | A0A2K5ZIM2 | 4 |
| PAPAN11413 | *Papio anubis* | A0A096NHY1 | 4 |
| COLAP30700 | *Colobus angolensis palliatus* | A0A2K5K1V9 | 4 |
| RHIBE34040 | *Rhinopithecus bieti* | A0A2K6MN44 | 4 |
| RHIRO31968 | *Rhinopithecus roxellana* | A0A2K6PBV8 | 4 |
| GORGO07778 | *Gorilla gorilla gorilla* | ENSGGOG00000003707 | 4 |
| HUMAN27992 | *Homo sapiens* | FUS_HUMAN | 4 |
| PANPA15069 | *Pan paniscus* | A0A2R9A818 | 4 |
| PANTR12223 | *Pan troglodytes* | H2RA80 | 4 |
| NOMLE17198 | *Nomascus leucogenys* | ENSNLEG00000016218 | 4 |
| AOTNA37715 | *Aotus nancymaae* | A0A2K5CEH5 | 4 |
| CALJA04844 | *Callithrix jacchus* | ENSCJAG00000003169 | 4 |
| SAIBB16827 | *Saimiri boliviensis boliviensis* | A0A2K6UY94 | 4 |
| CARSF03981 | *Carlito syrichta* | ENSTSYG00000011757 | 4 |
| PROCO05597 | *Propithecus coquereli* | A0A2K6G321 | 4 |
| BOVIN20650 | *Bos taurus* | A0A140T861 | 4 |
| CAPHI17070 | *Capra hircus* | A0A452DQG8 | 4 |
| PIGXX20573 | *Sus scrofa* | ENSSSCG00000030798.2 | 4 |
| VICPA09706 | *Vicugna pacos* | ENSVPAG00000004211 | 4 |
| CANLF16361 | *Canis lupus familiaris* | ENSCAFG00000016862 | 4 |
| VULVU24268 | *Vulpes vulpes* | ENSVVUG00000027797.1 | 4 |
| MUSPF04187 | *Mustela putorius furo* | M3YJ27 | 4 |
| AILME07018 | *Ailuropoda melanoleuca* | G1LKT4 | 4 |
| URSAM07884 | *Ursus americanus* | A0A452QX30 | 4 |
| FELCA17000 | *Felis catus* | ENSFCAG00000027928 | 4 |
| PTEVA04324 | *Pteropus vampyrus* | ENSPVAG00000003097 | 4 |
| MYOLU16715 | *Myotis lucifugus* | G1P397 | 4 |
| ERIEU03865 | *Erinaceus europaeus* | ENSEEUG00000004934 | 4 |

| | | | |
|---|---|---|---|
| EQUAS04518 | *Equus asinus* | ENSEASG00005011180.1 | 4 |
| MANJA23313 | *Manis javanica* | XP_017521732 | 4 |
| DASNO08858 | *Dasypus novemcinctus* | ENSDNOG00000017091 | 4 |
| CHOHO04231 | *Choloepus hoffmanni* | ENSCHOG00000010780 | 4 |
| SARHA12777 | *Sarcophilus harrisii* | G3WEM9 | 4 |
| MONDO11342 | *Monodelphis domestica* | ENSMODG00000001889 | 4 |
| MACEU03227 | *Macropus eugenii* | ENSMEUG00000000968 | 4 |
| PHACI08601 | *Phascolarctos cinereus* | ENSPCIG00000015897.1 | 4 |

## Nup153

| | | | |
|---|---|---|---|
| XENTR02157 | *Xenopus tropicalis* | ENSXETG00000014197 | 1 |
| CIOIN15442 | *Ciona intestinalis* | F6ZT40 | 1 |
| LEPOC02490 | *Lepisosteus oculatus* | W5MNS7 | 2 |
| ANGAN03227 | *Anguilla anguilla* | XP_035291407 | 2 |
| ANATE05308 | *Anabas testudineus* | A0A3Q1JXD7 | 2 |
| SERDU08610 | *Seriola dumerili* | A0A3B4U9G7 | 2 |
| CYNSE14136 | *Cynoglossus semilaevis* | A0A3P8WRB1 | 2 |
| SCOMX19706 | *Scophthalmus maximus* | ENSSMAG00000008418.1 | 2 |
| GASAC05602 | *Gasterosteus aculeatus* | G3NP66 | 2 |
| MOLML09099 | *Mola mola* | A0A3Q3X2J2 | 2 |
| TAKRU10012 | *Takifugu rubripes* | ENSTRUG00000010163 | 2 |
| TETNG09217 | *Tetraodon nigroviridis* | H3CUV8 | 2 |
| ORYLA02168 | *Oryzias latipes* | ENSORLG00000012674 | 2 |
| ORYME27144 | *Oryzias melastigma* | A0A3B3BZH3 | 2 |
| NOTFU28633 | *Nothobranchius furzeri* | XP_015829300 | 2 |
| KRYMA01537 | *Kryptolebias marmoratus* | A0A3Q3F4A6 | 2 |
| CYPVA16762 | *Cyprinodon variegatus* | A0A3Q2EGJ3 | 2 |
| POEFO03854 | *Poecilia formosa* | A0A087Y5I5 | 2 |
| POERE03600 | *Poecilia reticulata* | A0A3P9NMY4 | 2 |
| XIPMA17891 | *Xiphophorus maculatus* | ENSXMAG00000003437 | 2 |
| ASTCA22790 | *Astatotilapia calliptera* | A0A3P8Q5S6 | 2 |
| HAPBU04455 | *Haplochromis burtoni* | A0A3Q2V3R9 | 2 |
| NEOBR07702 | *Neolamprologus brichardi* | A0A3Q4H186 | 2 |
| ORENI21884 | *Oreochromis niloticus* | I3J2S0 | 2 |
| AMPCI22627 | *Amphilophus citrinellus* | A0A3Q0RT68 | 2 |
| AMPOC02144 | *Amphiprion ocellaris* | A0A3Q1ASF8 | 2 |
| AMPPE25635 | *Amphiprion percula* | A0A3P8SWR4 | 2 |
| HIPCM26771 | *Hippocampus comes* | A0A3Q2YQJ3 | 2 |
| GADMO09197 | *Gadus morhua* | ENSGMOG00000010422 | 2 |
| ESOLU15511 | *Esox lucius* | A0A3P9AD61 | 2 |
| ASTMX04317 | *Astyanax mexicanus* | ENSAMXG00000021025 | 2 |
| PYGNA00871 | *Pygocentrus nattereri* | A0A3B4D0Y3 | 2 |
| ICTPU18872 | *Ictalurus punctatus* | A0A2D0PV24 | 2 |
| DANRE15074 | *Danio rerio* | A0A0R4IVZ0 | 2 |
| LATCH05943 | *Latimeria chalumnae* | H3ACD6 | 2 |
| OCHPR04515 | *Ochotona princeps* | ENSOPRG00000006221 | 3 |
| COLAP01419 | *Colobus angolensis palliatus* | A0A2K5IW93 | 3 |
| CARSF08065 | *Carlito syrichta* | ENSTSYG00000001251 | 3 |
| URSMA09101 | *Ursus maritimus* | A0A452USF3 | 3 |
| ERIEU07856 | *Erinaceus europaeus* | ENSEEUG00000010183 | 3 |
| MELGA08705 | *Meleagris gallopavo* | G1MW17 | 3 |
| CHICK14841 | *Gallus gallus* | A0A1D5PXQ1 | 3 |
| FICAL02256 | *Ficedula albicollis* | U3KG17 | 3 |
| PARMJ14580 | *Parus major* | ENSPMJG00000018241.1 | 3 |
| JUNHY01031 | *Junco hyemalis* | ENSJHYG00000015545.1 | 3 |
| TAEGU05586 | *Taeniopygia guttata* | ENSTGUG00000005917 | 3 |
| SERCA00052 | *Serinus canaria* | ENSSCAG00000000585.1 | 3 |
| MELUD05600 | *Melopsittacus undulatus* | ENSMUNG00000009250.1 | 3 |
| CHRPI05922 | *Chrysemys picta bellii* | ENSCPBG00000008333.1 | 3 |
| CHEAB03264 | *Chelonoidis abingdonii* | ENSCABG00000005639.1 | 3 |
| PELSI16509 | *Pelodiscus sinensis* | K7GGR9 | 3 |
| SPHPU04687 | *Sphenodon punctatus* | ENSSPUG00000016687.1 | 3 |
| ANOCA05697 | *Anolis carolinensis* | G1KCK7 | 3 |
| ORNAN12721 | *Ornithorhynchus anatinus* | ENSOANG00000011858 | 4 |
| LOXAF08097 | *Loxodonta africana* | G3TFY7 | 4 |
| RABIT14444 | *Oryctolagus cuniculus* | G1SIZ4 | 4 |
| DIPOR15101 | *Dipodomys ordii* | ENSDORG00000001736 | 4 |
| HETGA17013 | *Heterocephalus glaber* | ENSHGLG00100001529.1 | 4 |
| CAVPO07965 | *Cavia porcellus* | ENSCPOG00000004007 | 4 |
| CHILA18633 | *Chinchilla lanigera* | ENSCLAG00000001915.1 | 4 |
| OCTDE06300 | *Octodon degus* | ENSODEG00000005620.1 | 4 |
| JACJA19740 | *Jaculus jaculus* | ENSJJAG00000016760.1 | 4 |
| CRIGR01418 | *Cricetulus griseus* | ENSCGRG00000009244.1 | 4 |
| MOUSE10788 | *Mus musculus* | E9Q3G8 | 4 |
| RATNO08849 | *Rattus norvegicus* | G3V662 | 4 |
| NANGA14919 | *Nannospalax galili* | ENSNGAG00000013817.1 | 4 |
| ICTTR03195 | *Ictidomys tridecemlineatus* | ENSSTOG00000005807 | 4 |
| CERAT09565 | *Cercocebus atys* | A0A2K5NJ22 | 4 |
| CHLSB06650 | *Chlorocebus sabaeus* | A0A0D9R5K0 | 4 |
| MACFA33008 | *Macaca fascicularis* | A0A2K5TL19 | 4 |
| MACMU16280 | *Macaca mulatta* | ENSMMUG00000001993 | 4 |
| MACNE10264 | *Macaca nemestrina* | A0A2K6CZZ2 | 4 |

| | | | |
|---|---|---|---|
| MANLE36126 | *Mandrillus leucophaeus* | A0A2K5YCI0 | 4 |
| RHIBE28287 | *Rhinopithecus bieti* | A0A2K6MPT9 | 4 |
| RHIRO11751 | *Rhinopithecus roxellana* | A0A2K6P3I1 | 4 |
| GORGO17159 | *Gorilla gorilla gorilla* | ENSGGOG00000014261 | 4 |
| HUMAN83246 | *Homo sapiens* | NU153_HUMAN | 4 |
| PANPA33599 | *Pan paniscus* | A0A2R9A8E2 | 4 |
| PANTR37299 | *Pan troglodytes* | A0A2I3SL61 | 4 |
| PONAB13047 | *Pongo abelii* | H2PI06 | 4 |
| NOMLE06341 | *Nomascus leucogenys* | G1QKG7 | 4 |
| AOTNA16236 | *Aotus nancymaae* | A0A2K5BV75 | 4 |
| CALJA14506 | *Callithrix jacchus* | F7EHG8 | 4 |
| SAIBB12360 | *Saimiri boliviensis boliviensis* | A0A2K6T874 | 4 |
| MICMU04339 | *Microcebus murinus* | ENSMICG00000006089 | 4 |
| PROCO19645 | *Propithecus coquereli* | A0A2K6EY30 | 4 |
| OTOGA13338 | *Otolemur garnettii* | H0XC65 | 4 |
| BOVIN19169 | *Bos taurus* | A0A3Q1MAS6 | 4 |
| CAPHI15081 | *Capra hircus* | A0A452FV25 | 4 |
| SHEEP10664 | *Ovis aries* | W5PPG7 | 4 |
| PIGXX30834 | *Sus scrofa* | A0A480MLT8 | 4 |
| VICPA01837 | *Vicugna pacos* | ENSVPAG00000002072 | 4 |
| CANLF13754 | *Canis lupus familiaris* | ENSCAFG00000010146 | 4 |
| VULVU13037 | *Vulpes vulpes* | A0A3Q7S734 | 4 |
| MUSPF09744 | *Mustela putorius furo* | ENSMPUG00000005159 | 4 |
| AILME04788 | *Ailuropoda melanoleuca* | G1L1N5 | 4 |
| FELCA05138 | *Felis catus* | ENSFCAG00000011583 | 4 |
| PTEVA10137 | *Pteropus vampyrus* | ENSPVAG00000012632 | 4 |
| MYOLU06869 | *Myotis lucifugus* | G1P376 | 4 |
| SORAR06050 | *Sorex araneus* | ENSSARG00000013269 | 4 |
| EQUAS08047 | *Equus asinus* | ENSEASG00005009945.1 | 4 |
| HORSE09415 | *Equus caballus* | ENSECAG00000016010 | 4 |
| MANJA39534 | *Manis javanica* | XP_036850301 | 4 |
| DASNO13918 | *Dasypus novemcinctus* | ENSDNOG00000001153 | 4 |
| MONDO04745 | *Monodelphis domestica* | ENSMODG00000010951 | 4 |
| PHACI10774 | *Phascolarctos cinereus* | ENSPCIG00000029680.1 | 4 |
| | **TNRC6A** | | |
| ANATE29102 | *Anabas testudineus* | A0A3Q1HCA4 | 2 |
| SERDU24303 | *Seriola dumerili* | A0A3B4UQH3 | 2 |
| CYNSE12687 | *Cynoglossus semilaevis* | A0A3P8WUF2 | 2 |
| SCOMX13694 | *Scophthalmus maximus* | ENSSMAG00000008293.1 | 2 |
| GASAC08536 | *Gasterosteus aculeatus* | G3PMS9 | 2 |
| MOLML21320 | *Mola mola* | A0A3Q3WUT0 | 2 |
| TAKRU11363 | *Takifugu rubripes* | ENSTRUG00000009839 | 2 |
| ORYLA16918 | *Oryzias latipes* | ENSORLG00000012798 | 2 |
| ORYME17187 | *Oryzias melastigma* | A0A3B3CNE4 | 2 |
| NOTFU05655 | *Nothobranchius furzeri* | A0A1A8A0C2 | 2 |
| KRYMA10889 | *Kryptolebias marmoratus* | A0A3Q3B4A0 | 2 |
| CYPVA18089 | *Cyprinodon variegatus* | A0A3Q2DHY7 | 2 |
| POEFO13649 | *Poecilia formosa* | A0A087XTC6 | 2 |
| POERE31547 | *Poecilia reticulata* | A0A3P9Q3C9 | 2 |
| XIPMA10115 | *Xiphophorus maculatus* | ENSXMAG00000001697 | 2 |
| ASTCA30898 | *Astatotilapia calliptera* | A0A3P8PDD9 | 2 |
| HAPBU25992 | *Haplochromis burtoni* | A0A3Q2WY78 | 2 |
| ORENI21771 | *Oreochromis niloticus* | I3KF62 | 2 |
| AMPCI16532 | *Amphilophus citrinellus* | A0A3Q0SBA3 | 2 |
| AMPOC21905 | *Amphiprion ocellaris* | A0A3Q1DGT8 | 2 |
| AMPPE09980 | *Amphiprion percula* | A0A3P8T1Z7 | 2 |
| HIPCM18813 | *Hippocampus comes* | A0A3Q2Z453 | 2 |
| ESOLU15775 | *Esox lucius* | ENSELUG00000013855.1 | 2 |
| ASTMX21398 | *Astyanax mexicanus* | ENSAMXG00000019919 | 2 |
| PYGNA21286 | *Pygocentrus nattereri* | A0A3B4DXJ0 | 2 |
| ICTPU25584 | *Ictalurus punctatus* | ENSIPUG00000001602.1 | 2 |
| DANRE29842 | *Danio rerio* | A0A0R4IJ26 | 2 |
| LATCH11830 | *Latimeria chalumnae* | H3ABQ4 | 2 |
| PROCA10536 | *Procavia capensis* | ENSPCAG00000008260 | 2 |
| OCHPR12848 | *Ochotona princeps* | ENSOPRG00000005819 | 2 |
| DIPOR07487 | *Dipodomys ordii* | ENSDORG00000014734 | 2 |
| ICTTR12692 | *Ictidomys tridecemlineatus* | ENSSTOG00000023394 | 2 |
| CARSF10784 | *Carlito syrichta* | ENSTSYG00000012711 | 2 |
| MICMU06378 | *Microcebus murinus* | ENSMICG00000001883 | 2 |
| CHOHO09049 | *Choloepus hoffmanni* | ENSCHOG00000009374 | 2 |
| MELGA04507 | *Meleagris gallopavo* | G1N3Z6 | 2 |
| PELSI06952 | *Pelodiscus sinensis* | ENSPSIG00000003965 | 2 |
| PAPAN11340 | *Papio anubis* | ENSPANG00000018523 | 3 |
| MUSPF02299 | *Mustela putorius furo* | M3YLM1 | 3 |
| SARHA15034 | *Sarcophilus harrisii* | G3W445 | 3 |
| MONDO11658 | *Monodelphis domestica* | ENSMODG00000016348 | 3 |
| ANAPL06809 | *Anas platyrhynchos* | ENSAPLG00000009978 | 3 |
| ANAPP26723 | *Anas platyrhynchos platyrhynchos* | A0A493TP73 | 3 |
| CHICK02910 | *Gallus gallus* | F1NAU7 | 3 |
| FICAL04956 | *Ficedula albicollis* | U3JMB2 | 3 |
| PARMJ04275 | *Parus major* | ENSPMJG00000019495.1 | 3 |

| | | | |
|---|---|---|---|
| JUNHY15710 | *Junco hyemalis* | ENSJHYG00000002230.1 | 3 |
| TAEGU02638 | *Taeniopygia guttata* | ENSTGUG00000006142 | 3 |
| SERCA11590 | *Serinus canaria* | ENSSCAG00000003490.1 | 3 |
| MELUD20429 | *Melopsittacus undulatus* | ENSMUNG00000002613.1 | 3 |
| CHRPI01613 | *Chrysemys picta bellii* | ENSCPBG00000001881.1 | 3 |
| CHEAB20009 | *Chelonoidis abingdonii* | ENSCABG00000016518.1 | 3 |
| SPHPU07887 | *Sphenodon punctatus* | ENSSPUG00000010525.1 | 3 |
| ANOCA11291 | *Anolis carolinensis* | ENSACAG00000011857 | 3 |
| XENTR01823 | *Xenopus tropicalis* | ENSXETG00000030437 | 3 |
| ORNAN00398 | *Ornithorhynchus anatinus* | ENSOANG00000007672 | 4 |
| LOXAF05468 | *Loxodonta africana* | G3STU5 | 4 |
| ECHTE12195 | *Echinops telfairi* | ENSETEG00000014658 | 4 |
| RABIT11241 | *Oryctolagus cuniculus* | ENSOCUG00000004325.3 | 4 |
| HETGA19908 | *Heterocephalus glaber* | ENSHGLG00100012496.1 | 4 |
| CHILA19632 | *Chinchilla lanigera* | ENSCLAG00000010217.1 | 4 |
| OCTDE00070 | *Octodon degus* | ENSODEG00000012802.1 | 4 |
| CRIGR22441 | *Cricetulus griseus* | ENSCGRG00000013351.1 | 4 |
| MOUSE56368 | *Mus musculus* | TNR6A_MOUSE | 4 |
| RATNO01902 | *Rattus norvegicus* | A0A0G2JZJ3 | 4 |
| NANGA18433 | *Nannospalax galili* | ENSNGAG00000020702.1 | 4 |
| CERAT31849 | *Cercocebus atys* | A0A2K5L032 | 4 |
| CHLSB14521 | *Chlorocebus sabaeus* | A0A0D9R2K7 | 4 |
| MACFA24538 | *Macaca fascicularis* | A0A2K5WC69 | 4 |
| MACMU14332 | *Macaca mulatta* | ENSMMUG00000020428 | 4 |
| MACNE37716 | *Macaca nemestrina* | A0A2K6CFH5 | 4 |
| MANLE09220 | *Mandrillus leucophaeus* | A0A2K6AAZ0 | 4 |
| COLAP38355 | *Colobus angolensis palliatus* | A0A2K5K3P5 | 4 |
| RHIBE00537 | *Rhinopithecus bieti* | A0A2K6JVI4 | 4 |
| GORGO07659 | *Gorilla gorilla gorilla* | ENSGGOG00000005137 | 4 |
| HUMAN27314 | *Homo sapiens* | TNR6A_HUMAN | 4 |
| PANPA14856 | *Pan paniscus* | A0A2R9A4V3 | 4 |
| PANTR11987 | *Pan troglodytes* | H2R4F2 | 4 |
| PONAB04808 | *Pongo abelii* | H2NQF8 | 4 |
| NOMLE05264 | *Nomascus leucogenys* | ENSNLEG00000011851 | 4 |
| AOTNA13844 | *Aotus nancymaae* | A0A2K5D1C1 | 4 |
| CALJA04720 | *Callithrix jacchus* | ENSCJAG00000020677 | 4 |
| SAIBB17092 | *Saimiri boliviensis boliviensis* | A0A2K6V8D8 | 4 |
| PROCO17866 | *Propithecus coquereli* | A0A2K6ENQ1 | 4 |
| OTOGA06942 | *Otolemur garnettii* | H0WUN8 | 4 |
| TUPBE12837 | *Tupaia belangeri* | ENSTBEG00000014965 | 4 |
| BOVIN20475 | *Bos taurus* | G3N258 | 4 |
| CAPHI16912 | *Capra hircus* | A0A452F9V6 | 4 |
| SHEEP12355 | *Ovis aries* | W5Q977 | 4 |
| PIGXX38979 | *Sus scrofa* | A0A287BNE0 | 4 |
| CANLF16466 | *Canis lupus familiaris* | ENSCAFG00000017531 | 4 |
| VULVU34014 | *Vulpes vulpes* | ENSVVUG00000005380.1 | 4 |
| AILME02472 | *Ailuropoda melanoleuca* | G1LRX8 | 4 |
| URSAM01166 | *Ursus americanus* | A0A452QQI8 | 4 |
| URSMA28208 | *Ursus maritimus* | A0A452TSR0 | 4 |
| PTEVA04038 | *Pteropus vampyrus* | ENSPVAG00000012066 | 4 |
| MYOLU03029 | *Myotis lucifugus* | G1NYH2 | 4 |
| ERIEU09113 | *Erinaceus europaeus* | ENSEEUG00000004530 | 4 |
| EQUAS29758 | *Equus asinus* | ENSEASG00005017693.1 | 4 |
| HORSE04775 | *Equus caballus* | ENSECAG00000025056 | 4 |
| MANJA36779 | *Manis javanica* | XP_036847877 | 4 |
| DASNO11072 | *Dasypus novemcinctus* | ENSDNOG00000013989 | 4 |
| PHACI20675 | *Phascolarctos cinereus* | ENSPCIG00000012106.1 | 4 |

[1]The taxonomic rank. 1: (Cordata; 2: (Vertebrata; 3: (Tetrapoda; 4: (Mammalia)))).

## References


1. Romero, Obradovic & Dunker, K. Sequence Data Analysis for Long Disordered Regions Prediction in the Calcineurin Family. *Genome Inform Ser Workshop Genome Inform* **8**, 110-124 (1997).
2. Radivojac, P., Obradovic, Z., Brown, C.J. & Dunker, A.K. Prediction of boundaries between intrinsically ordered and disordered protein regions. *Pac Symp Biocomput*, 216-27 (2003).
3. Hu, G. et al. flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nat Commun* **12**, 4438 (2021).
4. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat Rev Genet* **15**, 829-45 (2014).
5. Zhao, B. et al. Intrinsic Disorder in Human RNA-Binding Proteins. *J Mol Biol* **433**, 167229 (2021).
6. Altenhoff, A.M. et al. OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Res* **49**, D373-D379 (2021).
7. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539 (2011).
8. Cock, P.J. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-3 (2009).
9. Capra, J.A. & Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875-82 (2007).
10. Necci, M., Piovesan, D., Predictors, C., DisProt, C. & Tosatto, S.C.E. Critical assessment of protein intrinsic disorder prediction. *Nat Methods* **18**, 472-481 (2021).