Evolution of sequence traits of prion-like proteins linked to amyotrophic lateral sclerosis (ALS)

Jiayi Luo and Paul M. Harrison

Department of Biology, McGill University, Montreal, Quebec, Canada

ABSTRACT

Prions are proteinaceous particles that can propagate an alternative conformation to further copies of the same protein. They have been described in mammals, fungi, bacteria and archaea. Furthermore, across diverse organisms from bacteria to eukaryotes, prion-like proteins that have similar sequence characters are evident. Such prion-like proteins have been linked to pathomechanisms of amyotrophic lateral sclerosis (ALS) in humans, in particular TDP43, FUS, TAF15, EWSR1 and hnRNPA2. Because of the desire to study human disease-linked proteins in model organisms, and to gain insights into the functionally important parts of these proteins and how they have changed across hundreds of millions of years of evolution, we analyzed how the sequence traits of these five proteins have evolved across eukaryotes, including plants and metazoa. We discover that the RNA-binding domain architecture of these proteins is deeply conserved since their emergence. Prion-like regions are also deeply and widely conserved since the origination of the protein families for FUS, TAF15 and EWSR1, and since the last common ancestor of metazoa for TDP43 and hnRNPA2. Prion-like composition is uncommon or weak in any plant orthologs observed, however in TDP43 many plant proteins have equivalent regions rich in other amino acids (namely glycine and tyrosine and/or serine) that may be linked to stress granule recruitment. Deeply conserved low-complexity domains are identified that likely have functional significance.

BACKGROUND

Prions in eukaryotes have been linked to diseases, evolutionary capacitance, large-scale genetic control and long-term memory formation. Prion formation and propagation have been studied extensively, particularly in the model organism *Saccharomyces cerevisiae*, a budding yeast. *S. cerevisiae* has >200 prion-like proteins that tend to have N/Q-rich (asparagine/glutamine-rich) domains of the sort observed in >10 known prion-formers (*An, Fitzpatrick & Harrison, 2016*; *Harbi & Harrison, 2014*). In humans, prion-like proteins have been linked mechanistically to amyotrophic lateral sclerosis (ALS) and other neurological/neuromuscular disorders, in particular the RNA-binding proteins FUS, EWSR1, TAF15, TDP43 and hnRNPA2 (*Chen et al., 2019*; *Picchiarelli & Dupuis, 2020*; *Smethurst, Sidle & Hardy, 2015*). A schematic diagram of the domain structure of the
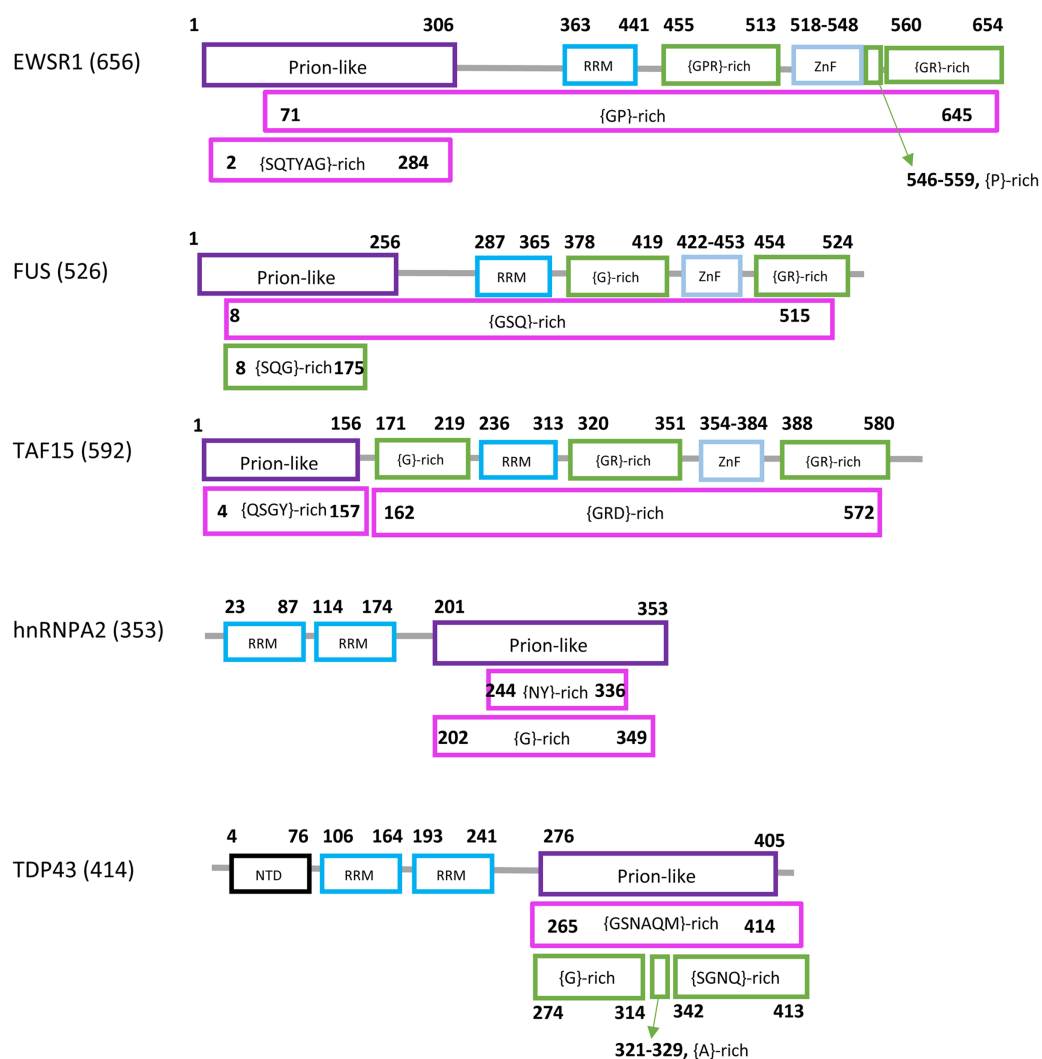
Figure 1 Domain content of the human forms of EWSR1, FUS, TAF15, hnRNPA2 and TDP43. The sequence names are at left with the sequence lengths in brackets. The sequences are represented by lines, with each domain as a box. The endpoints are labelled. Different types of domains are colour-coded. The endpoints of the prion-like regions (*purple*) are from PLAAC (*Lancaster et al., 2014*). The RRM (coloured *light blue*), ZnF (Zinc finger, *bright blue*) and NTD (N-terminal domain, *black*) assignments were determined as described in *Methods*. Compositionally-biased regions are labelled in the standard way according to fLPS output, {*xyz*}, meaning a bias for residues *x*, *y* and *z* in that order of precedence. They are coloured *magenta* if they were determined using default fLPS parameters except a threshold of $t = 1e-05$; they are coloured *green* if they are determined with fLPS parameters: $-m\ 5\ -M\ 25\ -t\ 1e-05\ -c$ *equal*. The latter parameter set is useful for labelling shorter 'low-complexity' regions. The data used to make this picture is available in tabular format in File S6. Full-size ◢ DOI: 10.7717/peerj.14417/fig-1

human forms of these proteins is illustrated (Fig. 1). Each has a prion-like region and at least one RRM RNA-binding domain.

FUS is an RNA-binding protein involved in transcriptional activation and DNA repair (*Aman et al., 1996*; *Law, Cann & Hicks, 2006*; *Naumann et al., 2018*; *Wang et al., 2013*). Mutations in FUS are associated with ~5% of inherited ALS cases (*Monahan et al., 2018*); some of these mutations may cause intranuclear aggregation of FUS as part of its role in

pathomechanisms (*Nomura et al., 2014*). TDP43 has been shown to have multiple roles in repression of transcription, alternative splicing regulation and translational regulation (*Mitra et al., 2019*; *Ou et al., 1995*; *Sephton et al., 2011*). TDP43 is found in neuronal cytoplasmic aggregates in most ALS patients, although only a small minority of these have pathogenic mutations in the TDP43 gene (*Mackenzie et al., 2007*; *Sreedharan et al., 2008*). TDP43 has two close relatives, EWSR1 and TAF15. EWSR1 functions as a transcriptional activator (*Ohno, Rao & Reddy, 1993*), and is best known from its role in forming a chimeric oncoprotein linked to Ewing sarcoma and other tumors (*Anderson et al., 2018*). TAF15 is a TATA-binding protein that is a subunit of transcription factor IID (*Bertolotti et al., 1998*). TAF15 and EWSR1 have also been linked to neuronal aggregation and pathogenic mutations (*Couthouis et al., 2012*; *Jackrel & Shorter, 2014*). hnRNPA2/B1 functions diversely in mRNA transport, processing and metabolism (*Kim et al., 2013*; *Zhao et al., 2018*). It has been also found in neuronal cytoplasmic aggregates linked to ALS, and bears pathogenic mutations that are linked to both ALS and multi-system proteinopathy (*Kim et al., 2013*; *Zhao et al., 2018*).

Experiments using mutated sequences have demonstrated that the prion-like regions, low-complexity domains and intrinsically disordered parts of hnRNPA2, FUS and TDP43 proteins are linked to protein aggregation and phase separation into membraneless organelles, in particular stress granules (*Kim et al., 2013*; *Molliex et al., 2015*; *Patel et al., 2015*; *Wang et al., 2018*). *Molliex et al. (2015)* showed that low-complexity domains of hnRNPA2 mediate liquid-liquid phase separation and stress granule assembly. *Patel et al. (2015)* demonstrated that the physiological role of FUS depends on liquid-liquid phase separation and that disease-linked mutation promotes a liquid to solid phase transition. Different types of amino-acid residues appear to have distinct biophysical roles, for example in the liquid-liquid phase separation of FUS, tyrosine and arginine residues control the saturation concentration of phase separation, glycine enhances liquidity, and glutamine and serine promote the solidification of aggregates (*Wang et al., 2018*). In a recent study, it was observed that the tyrosine residues in the prion-like region in FUS are amongst the absolutely conserved residues across eukaryotes, which include also some of the glutamine, serine and glycine residues (*Dasmeh & Wagner, 2021*). TDP-43 lacks the arginine and glycine content seen in FUS and hnRNPA2, and thus the molecular determinants of phase separation differ; within the TDP-43 prion-like region there is a conserved tract that forms an alpha-helix that is important for phase separation in isolated prion-like subsequences (*Conicella et al., 2016*). Aromatic residues adjacent to glycines and serines also contribute to TDP-43 phase separation *in vitro* (*Li et al., 2018*). Many experiments have sought to use the budding yeast *Saccharomyces cerevisiae* as a model system for studying the aggregation of these proteins, or the aggregation of homologous or compositionally similar domains (*Monahan et al., 2018*).

Here, we have probed into the evolution across eukaryotes of the sequence traits of these proteins that are linked to ALS in humans, namely TDP43, FUS, TAF15, EWSR1 and hnRNPA2. Evolutionary trends are examined chiefly across plants and metazoans, since most of the data is available from these kingdoms. We discover that the RNA-binding domain architecture of these proteins is deeply conserved since their origination.

Prion-like regions are also prevalent since the origination of the protein families for FUS, TAF15 and EWSR1, and since the last common ancestor of metazoans for TDP43 and hnRNPA2. Prion-like composition is rare in any plant orthologs observed, however in TDP43 many plant proteins have equivalent regions rich in glycine and tyrosine and/or arginine, that may be linked to stress granule recruitment. Specific compositional biases predominate in many cases for these proteins across clades as ancient as the vertebrates, or even the metazoa. Shorter low-complexity regions can also be deeply conserved across kingdoms, indicating functional significance.

METHODS

Data

Sets of orthologs for the following protein families were downloaded from OrthoDB database version 10.1 at the Eukaryota level (*Kriventseva et al., 2019*): TDP43 (Group 1425837at2759), hnRNPA2 (Group 1202220at2759) and EWS/TAF15/FUS (Group 1539664at2759). These data also contain paralogs. For some analyses of the hnRNPA2 data, only proteins from a reduced list of representative eukaryotes used in a previous investigation were included (*An & Harrison, 2016*; *Su & Harrison, 2020*).

Multiple sequence alignment and phylogenetic trees

Multiple sequence alignments were constructed using Clustal Omega (*Sievers & Higgins, 2018*), using default parameters. Phylogenetic trees were made using PhyML with Bayesian information criterion and aBayes branch support, and other parameters at defaults (*Guindon et al., 2005*). The aBayes branch support is a Bayesian-like transformation of the standard approximate likelihood ratio test implemented in PhyML, that is very fast and has high statistical power (*Anisimova et al., 2011*). Trees were saved in Newick format for input into the phylogenetic tree drawing tool Evolview version 2 (*Subramanian et al., 2019*). Annotations for prion-like regions, structured protein domains and compositionally biased regions (annotation described below) were formatted for input into Evolview using AWK scripts.

Annotation of protein domains

From InterPro version 66.0, we extracted Pfam (version 34.0) protein domain positions (*El-Gebali et al., 2019*; *Mitchell et al., 2019*). These were checked against PROSITE and SMART domain annotations (*Letunic & Bork, 2018*; *Sigrist et al., 2010*), as in a previous study (*Su & Harrison, 2020*), and they diverge by <1% in the total number of domains labelled. Since there were some OrthoDB sequences that do not have Pfam annotations, we ran the Hmmsearch program using domain Hidden Markov Models (HMMs) downloaded from the Pfam website (http://pfam.xfam.org/) to label further domains (*El-Gebali et al., 2019*; *Mitchell et al., 2019*), with e-value threshold 0.01. Also, some further RRM RNA-binding domain annotations were made by searching for RRM domain protein structures in the PDB (Protein Data Bank) (*Bittrich et al., 2021*), and extracting the relevant protein domain sequences from the ASTRALSCOP database version 2.06 using PDB identifiers and residue ranges (*Fox, Brenner & Chandonia, 2014*). These sequences

were then compared to the proteomes using BLASTP 2.9.0 (*Altschul et al., 1997*) with e-value threshold 1e−04. All these annotations were then reduced for overlap by giving HMM-derived annotations precedence and otherwise sorting them in decreasing order of domain length and progressively flagging overlappers further down the lists for deletion.

Prion-like regions, intrinsic disorder, and predicted stress-granule recruitment

Prion-like regions were annotated using the PLAAC program with default parameters (*Lancaster et al., 2014*). Experiments in *S. cerevisiae* wherein prion-forming sequences are scrambled yet maintain their prion nature have shown that prion-forming domains can be compositionally defined (*Ross, Baxa & Wickner, 2004*). PLAAC uses a Hidden Markov Model trained on known prion-forming domains from budding yeast that were detected through several tests for prion-like behaviour and amyloid formation (*Alberti et al., 2009*). Prion-like regions have compositions judged similar to those that can form prions. Although prion-like composition is typically identified as based on glutamine and asparagine bias, often regions that are biased for subsidiary residues of prion-like regions obtain high prion-like composition scores, *e.g.*, regions rich in tyrosine, glycine and serine. Prion-like regions were labelled if the PLAAC PRD prion-like composition score was >0.0 (as in a previous analysis (*Su & Harrison, 2019*)), and colour-coded according to PRD score in the annotated trees. Higher scores >15.0 are observed for known prion-forming domains (*Su & Harrison, 2019*). The program fLPS2 was used to annotate compositionally-biased (CB) regions (*Harrison, 2006*; *Harrison, 2017*; *Harrison, 2021*). Expected background frequencies = 0.05 were used in running fLPS2, with a threshold P-value of 10^{-6}. CB regions were grouped together according to whether their bias signatures had a common form *{Xyzr….}*, where *X* is the main biasing residue and *yzr…*, *etc.* are subsidiary biases that can be in any order (*i.e.*, allowing for permutation of any subsidiary biases). The ten most common such CB regions are labelled on each phylogenetic tree. To analyze milder biases of experimental relevance, the '-z thorough' setting was used in fLPS2, with a threshold $P = 1 \times 10^{-4}$ (*Harrison, 2021*).

Total proportions of intrinsic disorder were calculated using IUPRED3 with default parameters (*Dosztanyi, 2018*; *Ward et al., 2004*), as in a previous work (*Su & Harrison, 2020*).

The server SGNN was used to estimate whether sequences can be recruited to stress granules (*Iglesias et al., 2021*). This tool was derived through analysis of data for the budding yeast *S. cerevisiae*, and thus may miss some of the compositional characteristics that drive stress granule recruitment in other organisms. Also, it may not work for regions of proteins that do not have prion-like characteristics, since it was trained to identify prion-like regions that are stress-granule recruited.

RESULTS

Phylogenetic trees were calculated for eukaryotic protein families of TDP43, hnRNPA2 and EWSR1/TAF15/FUS. Protein domain content, prion-like composition, compositional biases, low-complexity regions, intrinsic disorder, sequence length and predicted stress

granule membership were calculated and annotated onto the trees (File S1). The data for making and annotating these trees is provided in File S2. Results are presented as follows:

  i)   trends in conservation of features since the last common ancestor of eukaryotes;
 ii)   distinct protein forms that are discovered in the metazoan and plant kingdoms;
iii)   phylogenetic origins of FUS, TAF15 and EWSR1;
iv)   specific evolutionary trends of sequence features for FUS, TAF15 and EWSR1.
 v)   trends in occurrences of shorter low-complexity regions (LCRs).

Deep conservation of protein domains and prion-like regions since last common ancestor of eukaryotes

RNA-binding protein domain content is deeply conserved since the last common ancestor of eukaryotes for each of the protein families examined. For the hnRNPA2 family, the same number of RRM RNA-binding domains is largely maintained since the last common ancestor (LCA) of eukaryotes (>88% of orthologs) (File S1; Fig. 2C). Across the TDP43 family, all plant species have an equivalent RRM domain where metazoans have the NTD (N-terminal domain) (File S1). The total number of domains in TDP43 (N-terminal domain and RRM RNA-binding domains) is also deeply conserved (over >88% of sequences) (Fig. 2A). This NTD N-terminal domain has been shown to potentially have either DNA- or RNA-binding ability (*Chang et al., 2012*; *Qin et al., 2014*). Over the EWSR1/TAF15/FUS tree, a single RRM domain is maintained across 95% of sequences (Fig. 2B).

The prion-like regions (as defined by PLAAC) are generally conserved since the LCA of metazoans for all three protein families studied (File S1), with the percentage of metazoan orthologs maintaining some prion-like composition 74% for TDP43, 79% for hnRNPA2 and 90% for EWSR1/TAF15/FUS. The most notable exceptions are for TDP43 in some specific metazoan clades (File S1).

Distinct protein forms in metazoans and plants

For the protein families that have substantial involvement in both plants and metazoans (*i.e.*, TDP43 and hnRNPA2), we can examine the distinct sequence characteristics in each of these kingdoms. These are summarized in Fig. 3. Most notably, the orthologs in plants have less intrinsic disorder and fewer prion-like regions with weaker prion-like compositions as defined using the PLAAC program (details in Fig. 2 and File S1). Interestingly however, the plant proteins are predicted to be more recruited to stress granules by the SGNN tool (details in Table 1). Zoom-ins of the plant and vertebrate parts of the TDP43 tree are shown in Fig. 4, to highlight these trends.

For TDP43 specifically, plant sequences do not at all assign to the NTD domain, but instead tend to a RRM domain assignment at the equivalent sequence position (File S1). Although they generally lack a prion-like region (with just a handful of exceptions; File S1), many of them (28%, 43/151) have a {GRY}-/{GY}-/{GS}-biased region instead at the same relative position in their sequences, and all but one of these is predicted to be recruited to
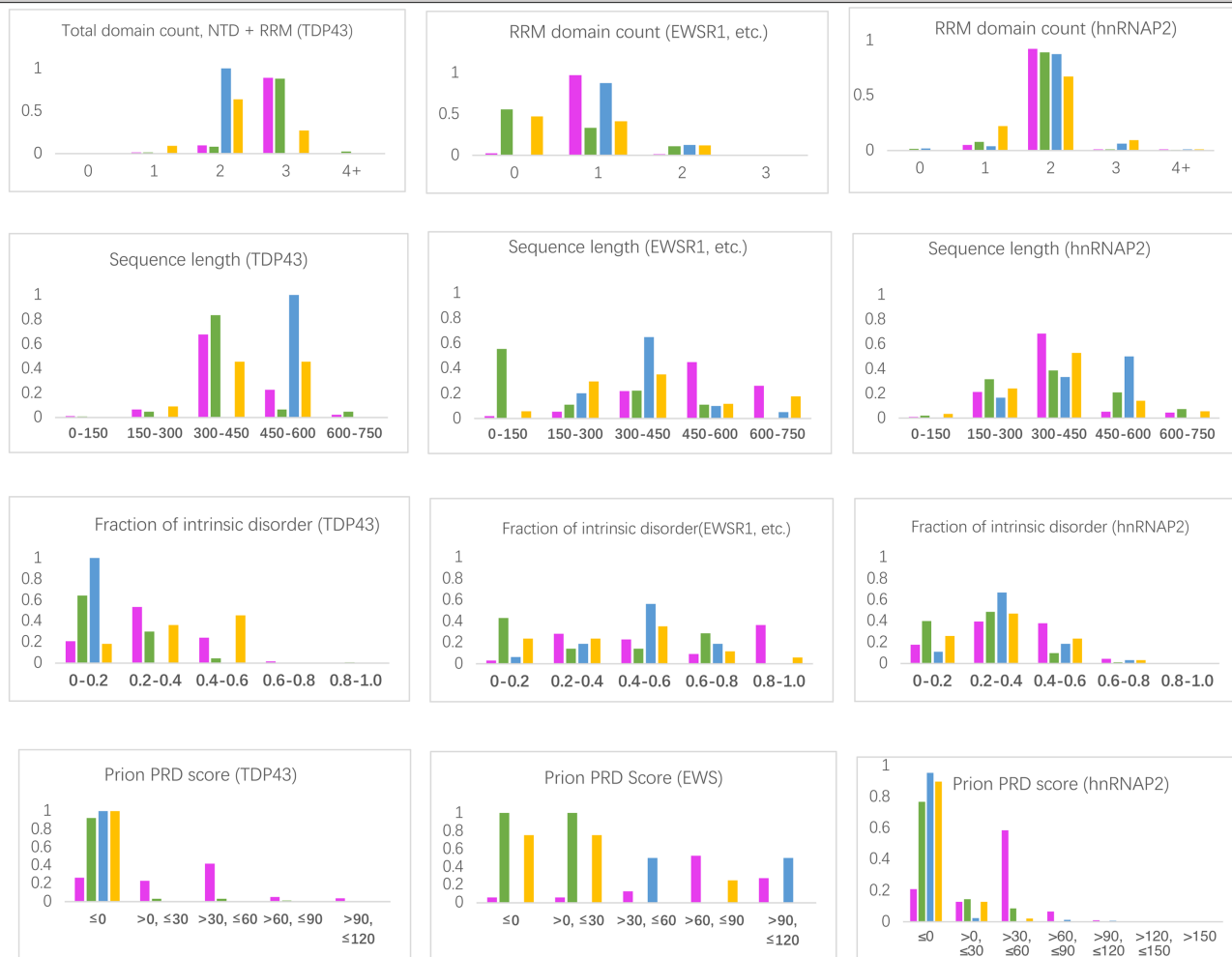
Figure 2 Bar charts for the three sets of orthologs analyzed for various attributes. They are arrayed in columns (A–C) for (total protein in brackets): TDP43 (763), EWSR1/TAF15/FUS (979), and hnRNPA2 (5503). The taxonomic groups are coloured as follows: magenta for metazoan, green for plants, blue for fungi and grey for others. The attributes analyzed (from top to bottom in each column): (i) *RRM RNA-binding domain count* for EWSR1/TAF15/FUS and total domain count for TDP43; (ii) *sequence length;* (iii) *proportion of intrinsic disorder*, as determined using the program IUPred3 (*Erdos, Pajkos & Dosztanyi, 2021*); (iv) *prion-like composition score* (PRD score) from the program PLAAC (*Lancaster et al., 2014*) for labelling regions of prion-like composition.   Full-size DOI: 10.7717/peerj.14417/fig-2

stress granules by the SGNN tool (*Iglesias et al., 2021*). A {GMNQS}-rich prion-like region is predominant across vertebrates, with the exception of some fish clades (File S1, region of labelled by star symbols). Other animals demonstrate diverse bias patterns that are conserved in a clade-specific manner. One prominent feature of the zoom-ins of both the plant and vertebrate tree areas is a short ~30-residue {G}- or {G*x*}-rich low-complexity region within the annotated prion-like regions, indicating a possible functional significance for such a domain (Fig. 4).

Similar trends are observed in the hnRNPA2 tree (File S1). A {GNY}-, {GNSY}-, or {GY}-rich prion-like region predominates in metazoans. However, few biased regions are

Figure 3 **Summary of the trends for TDP43 and hnRNPA2 comparing plants to metazoa.** The total numbers of sequences are labelled at the top of each section. The first five attributes are listed as mean ± standard deviation, except the last one, which has a percentage of the total number of sequences in brackets. The attributes are: (i) count of RRM RNA-binding domains for hnRNPA2 and total domains for TDP43 (N-terminal domain plus RRM domains); (ii) LLR (log-likelihood) score and PRD score from the PLAAC program (*Lancaster et al., 2014*); (iii) proportion of intrinsic disorder, as judged by IUPred3 (*Erdos, Pajkos & Dosztanyi, 2021*); (iv) sequence length; (v) count of positive predictions of stress granule membership by SGNN (*Iglesias et al., 2021*). The values are colour-coded according to the key used in colour-coding the tree annotations in File S1.   Full-size 🖼 DOI: 10.7717/peerj.14417/fig-3

observed in plant orthologs where they lack a prion-like region, and they maintain low levels of annotated intrinsic disorder (File S1; Fig. 3).

Origins of FUS, TAF15 and EWSR1

FUS likely became widely conserved in an ancestor of the *Bilateria*, i.e., animals with embryonic bilateral symmetry, although there are six orthologs in cnidarians, so the exact evolutionary timepoint of its emergence is difficult to discern, since there may have been some initial complex patterns of gene loss at such early evolutionary stages. In a similar manner, TAF15 and EWSR1 seem to have become conserved in an early ancestor of vertebrates, but there are a small number of diverse assigned orthologs outside of the vertebrate clade (18 non-vertebrates out of 250 total for TAF15, and 6 out of 280 for EWSR1), so here again the exact evolutionary timepoint of the emergence of these proteins

Table 1 Predictions of stress granule membership by SGNN.

| Protein family	Range (Number of proteins)	SGNN prediction of stress granule membership	
		Yes	No
EWSR1	Eukaryota (280)	4.6%	95.4%
FUS	Eukaryota (396)	13.6%	86.4%
TAF15	Eukaryota (250)	21.6%	78.4%
TDP43	Eukaryota (763)	81.7%	18.3%
TDP43	Metazoa (587)	79.9%	20.1%
TDP43	Plants (151)	87.4%	12.6%
TDP43	Fungi & Other (12)	75.0%	25.0%
hnRNPA2	Eukaryota (5,503)	38.9%	61.1%
hnRNPA2	Metazoa (2,338)	23.0%	77.0%
hnRNPA2	Plants (1,920)	63.0%	37.0%
hnRNPA2	Fungi (887)	22.4%	77.5%
hnRNPA2	Other (410)	52.9%	47.1%
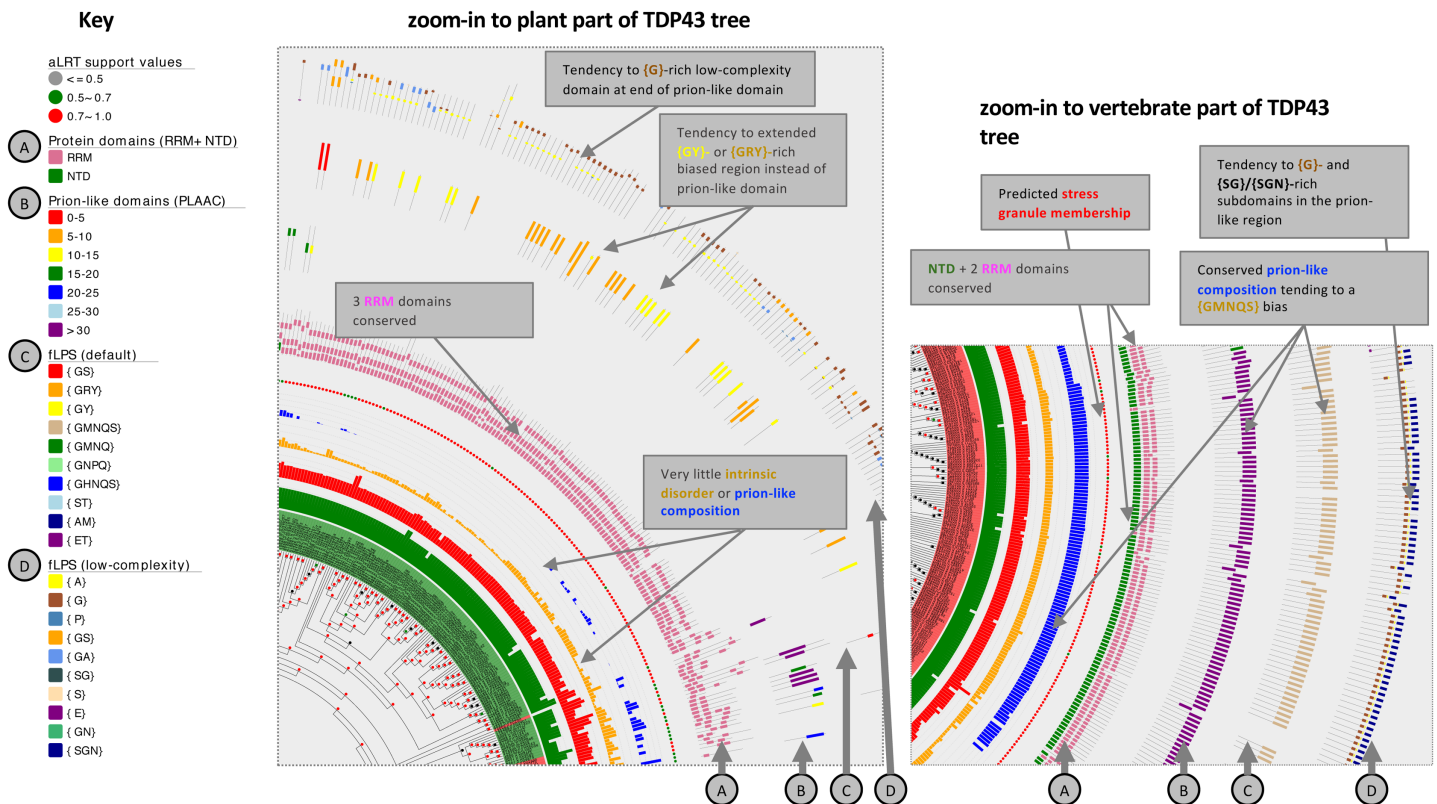
Figure 4 Zoom-ins of the TDP43 phylogenetic tree illustrating the detail of the plant and vertebrate ortholog annotations. The key for tree attributes is at left and is described in detail in File S1. The trees are drawn with Evolview version 2 (*Subramanian et al., 2019*). Salient distinguishing features of either tree zoom-in are labelled.
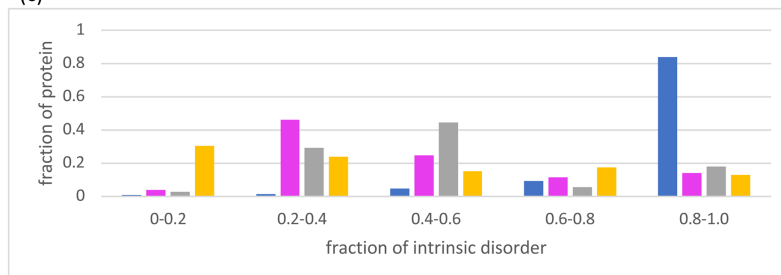Full-size ▣ DOI: 10.7717/peerj.14417/fig-4
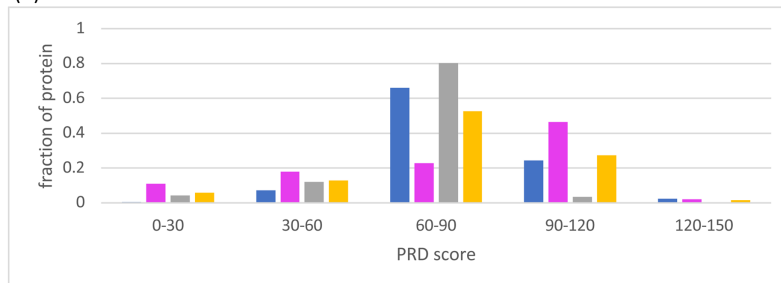
(a)
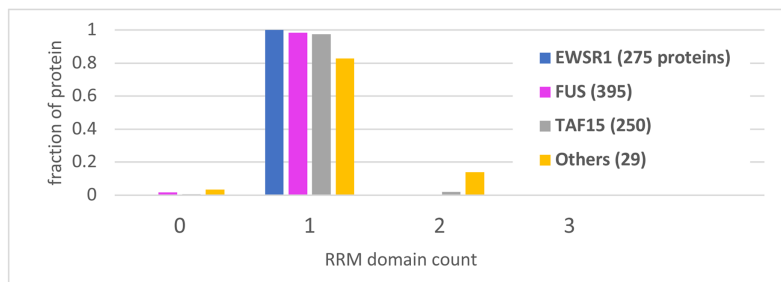
(b)
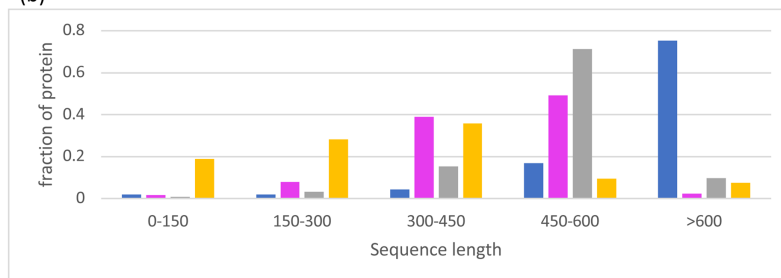
(c)

(d)

Figure 5 Bar charts for the tree of EWSR1/TAF15/FUS orthologs showing the distributions of various attributes for each of the proteins EWSR1, TAF15 and FUS. The total number of sequences for each of the proteins (plus 'others') are in the key at the top of the figure. The y-axis is the fraction of proportion of proteins. The figures panels are for: (A) RRM RNA-binding domain count; (B) sequence length; (C) proportion of intrinsic disorder, as determined using the program IUPred3 (*Erdos, Pajkos & Dosztanyi, 2021*); (D) PRD score from the program PLAAC (*Lancaster et al., 2014*) for labelling regions of prion-like composition.   Full-size 🖼 DOI: 10.7717/peerj.14417/fig-5

is also difficult to discern. Generally, the small number of non-metazoan homologs in the tree are much shorter sequences (particularly in plants), and correspondingly have less intrinsic disorder (Fig. 5).

Table 2 Zinc finger domains in the EWSR1/TAF15/FUS family.

Protein family	Number of Zn-finger domains			
	0	1	2	3+
EWSR1	27 (10%)	193 (69%)	59 (21%)	0
FUS	27 (7%)	349 (88%)	21 (5%)	0
TAF15	12 (5%)	235 (94%)	2 (1%)	0

Specific evolutionary trends for EWSR1, TAF15 and FUS

Apart from the deep conservation of RNA-binding domains and prion-like regions (Figs. 2 and 5), the individual sequence evolution of EWSR1, TAF15 and FUS has unfolded idiosyncratically.

In particular, each of these protein families have characteristic amounts of deeply conserved intrinsic disorder (Fig. 5C). The EWSR1 family has the highest proportions of annotated intrinsic disorder, typically >0.8, that have been conserved across clades since the last common ancestral sequence (Fig. 5C, File S1). Comparatively, TAF15 has intermediate levels of intrinsic disorder (mode 0.4–0.6), with the lowest for FUS proteins (mode 0.2–0.4) (Fig. 5C). These tendencies do not correlate with PLAAC prion-like composition scores, which are notably higher for FUS compared to the other two proteins (Fig. 5D). All three families maintain some degree of prion-like compositions (score >0) over the vast majority of the family members (97% of members for EWSR1, 93% for TAF15 and 85% for FUS).

A single Zn-finger domain is a deeply conserved component of both the FUS and TAF15 protein families (File S1; Table 2). However, EWSR1 has greater variation in the number of Zinc-finger domains compared to the other two protein families (Table 2), with one in five orthologs having two domains, and one in 10 having none.

The default fLPS program was used to detect longer compositionally biased regions (CBRs). We discovered that CBRs are maintained across deep and diverse clades when we examined the detail of the annotated trees (File S1). TAF15 conserves a {GR}- or {GY}-rich bias across mammals (bias P-values $< 1 \times 10^{-10}$) with the C-terminal R bias maintained in other vertebrates (labelled as a {CR} bias encompassing also the Zn-finger domains, P-values $< 1 \times 10^{-6}$). Where a predominant {GY} bias is labelled, it stretches across the whole sequence, whereas the {GR} bias is beyond the Zn-finger domains at the C-terminus. For FUS, the fLPS2 program assigns a biased region across most of the sequence that varies between {GQSY} or {GQRSY} in a clade-specific manner (bias P-value $< 1 \times 10^{-24}$). In EWSR1, a very specific {GMPRSTY} bias predominates across all vertebrates (bias P-value $< 1 \times 10^{-52}$), with the exception of fish, which conserve a similar, but also very specific {GMPQRSTY} bias (P-value $< 1 \times 10^{-46}$) (File S1).

We also examined for milder compositional biases of experimental relevance, i.e., involved in recruitment to stress granules, such as {R} and {Y} bias (fLPS2 P-value $\leq 1 \times 10^{-4}$) (Wang et al., 2018). Both {R}- and {Y}-biased regions occur with high frequency, but also other positively-charged regions, i.e., {K}-biased regions, are quite common, particularly in TAF15 orthologs (File S3). {R}-biased regions are characteristically shorter

than {Y}-biased regions (File S4), indicating that arginine residues corresponding to tyrosine residues that might interact with them are less dispersed along the sequence. The most prevalent bias across all protein types is {G}.

Low-complexity regions

We analyzed the distribution of low-complexity regions (LCRs) annotated using the program fLPS 2.0 and parameters recommended for shorter LCR discovery (*Harrison, 2021*). LCRs associated with RNA-binding, such as the {GR}-/{GPR}-/{G}-rich regions seen in EWSR, TAF15 and FUS, are sometimes termed RGG or RG regions, and associated functionally with RNA-binding (*Chong, Vernon & Forman-Kay, 2018*). In addition to these regions, we discovered that other LCRs are also deeply conserved across kingdoms for each of three protein families examined (File S5). For example, a short {P}-rich tract (as exemplified in EWSR1, Fig. 1) is deeply conserved across EWSR1 orthologs (219/249 cases) and some other non-metazoan orthologs (20/249 cases) (14.3 ± 5.4 residues in length, median *P*-value = 8.6e−08). A short {A}-rich LCR is deeply conserved across many clades for the TDP43 family, specifically within the putative prion-like domain in vertebrates (11.7 ± 8.9 residues in length, median *P*-value = 1.1e−06, visible in Fig. 4), and corresponds to an alpha-helical region that likely functions in phase separation (*Conicella et al., 2016*). In general, the top ten LCRs are conserved across kingdoms with few exceptions (non-grey bars in the charts, File S5).

DISCUSSION

Since they emerged evolutionarily, the proteins of these five families have had deeply conserved RNA-binding capability, that varies only rarely in terms of RNA-binding domain number, indicating selection pressure against further RNA-binding domain duplication. In tandem across metazoa, the TDP43 and hnRNPA2 families demonstrate a selection pressure to maintain prion-like composition, in contrast to the situation in the plants, where prion-like character is markedly rarer, particularly for TDP43 plant orthologs. However, such proteins may still function in biomolecular condensate formation, particularly plant stress-granule formation in response to environmental cues (*Cuevas-Velazquez & Dinneny, 2018*; *Maruri-Lopez et al., 2021*). The most common compositional biases in plant TDP43 orthologs are a group of {GRY}/{GY}/{GS} biases that may be linked to stress granule recruitment, indeed these proteins are almost universally predicted to be recruited to stress granules. Such subsidiary biases for G, Y or S are sometimes observed in prion-forming domains (*Lancaster et al., 2014*). Mutational experiments on FUS have shown that separate regions of arginine and tyrosine bias interact with each other through specific side-chain arginine-tyrosine bonding to control the saturation concentration at which liquid-liquid phase separation occurs, with glycine residues shown to be necessary to maintain membraneless organelle liquidity (*Portz, Lee & Shorter, 2021*; *Wang et al., 2018*). Tyrosine-tyrosine interactions also contribute to phase separation (*Sun et al., 2011*). In general, however the five protein families examined here do not demonstrate the very deep eukaryote-wide conservation of both RNA-binding domain architecture and prion-like composition observed for the TIA-1 protein family

(*An & Harrison, 2016*; *Su & Harrison, 2020*). TIA-1 has been shown to form prion-like protein aggregates in both its yeast and human forms and is also linked mechanistically to ALS, and through rare inherited mutations in some human cohorts (*Gu et al., 2018*; *Zhang et al., 2018*).

A caveat here is that the SGNN tool used to estimate stress granule recruitment was only trained on *S. cerevisiae* proteins with prion-like composition, so that it is unclear what results for proteins with weak prion-like composition, especially in divergent eukaryotes, might mean. Also, it is important to bear in mind that the PLAAC program for identification of prion-like composition was trained on data obtained from experiments on *S. cerevisiae* proteins. However, there has been some success in using the PLAAC program to identify prion-forming domains in other eukaryotes, and in bacteria and archaea (*Chakrabortee et al., 2016*; *Harrison, 2019*; *Kim et al., 2013*; *Sideri et al., 2017*; *Yuan & Hochschild, 2017*; *Zajkowski et al., 2021*), but there may be other prion-forming domain compositions that are not sampled during the evolution of budding yeasts (*An, Fitzpatrick & Harrison, 2016*; *Wang & Harrison, 2021*).

In general, compositional biases in these protein families vary diversely from clade to clade, but can be deeply conserved within a clade, *e.g.*, a specific {GMNQS} bias conserved in TDP43 across vertebrates (with the exception of fish). This mode of evolution suggests periodic rare shifts in the molecular grammar governing the functional traits of these sequences, possibly linked to encoding of prion formation or stress granule recruitment. The maintenance of prion-like character despite underlying shifts in compositional bias is also a feature of the evolution of prion-forming domains of the budding yeast *Saccharomyces cerevisiae* (*Su & Harrison, 2019*). These ALS-linked ortholog sequences could thus be further analyzed mutationally to dissect how the side-chain interactions governing possible prion formation or stress granule recruitment capability may have changed at key time points of eukaryotic organismal evolution. Shorter low-complexity tracts that are conserved across diverse eukaryotic clades (such as the {P}-rich example that was highlighted) could also be dissected to discern their functional significance.

CONCLUSIONS

We have presented a detailed analysis of the evolution of sequence traits of five ALS-linked protein families. We discovered that the RNA-binding domain architecture of these proteins is deeply conserved since the origination of these proteins, or since the last common ancestor of metazoa. Prion-like regions are also prevalent since their origination for FUS, TAF15 and EWSR1, and since the last common ancestor of metazoa for TDP43 and hnRNPA2. We discussed how the plant orthologs of TDP43 and hnRNPA2 are distinct from their metazoan counterparts, but still often have features that can be linked to stress granule formation. These data are useful in developing further hypotheses about these ALS-linked protein families that can be experimentally tested, particularly when assessing how corresponding protein sequences have changed in model organisms. Specific conserved subdomains and low-complexity regions that were observed could be examined experimentally for their functional significance. This study demonstrates the

utility of combined application of diverse sequence annotation programs to characterize evolutionary trends.

ADDITIONAL INFORMATION AND DECLARATIONS

Competing Interests
The authors declare that they have no competing interests.

Author Contributions
- Jiayi Luo performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Paul M. Harrison conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability
The following information was supplied regarding data availability:
The raw data is available in the Supplemental Files.

Supplemental Information
Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.14417#supplemental-information.

REFERENCES
Alberti S, Halfmann R, King O, Kapila A, Lindquist S. 2009. A systematic survey identifies prions and illuminates sequence features of prionogenic proteins. *Cell* **137(1)**:146–158 DOI 10.1016/j.cell.2009.02.044.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25(17)**:3389–3402 DOI 10.1093/nar/25.17.3389.

Aman P, Panagopoulos I, Lassen C, Fioretos T, Mencinger M, Toresson H, Hoglund M, Forster A, Rabbitts TH, Ron D, Mandahl N, Mitelman F. 1996. Expression patterns of the human sarcoma-associated genes FUS and EWS and the genomic structure of FUS. *Genomics* **37(1)**:1–8 DOI 10.1006/geno.1996.0513.

An L, Fitzpatrick D, Harrison PM. 2016. Emergence and evolution of yeast prion and prion-like proteins. *Bmc Evolutionary Biology* **16(1)**:24 DOI 10.1186/s12862-016-0594-3.

An L, Harrison PM. 2016. The evolutionary scope and neurological disease linkage of yeast-prion-like proteins in humans. *Biology Direct* **11(1)**:32 DOI 10.1186/s13062-016-0134-5.

Anderson ND, de Borja R, Young MD, Fuligni F, Rosic A, Roberts ND, Hajjar S, Layeghifard M, Novokmet A, Kowalski PE, Anaka M, Davidson S, Zarrei M, Id Said B, Schreiner LC, Marchand R, Sitter J, Gokgoz N, Brunga L, Graham GT, Fullam A, Pillay N, Toretsky JA, Yoshida A, Shibata T, Metzler M, Somers GR, Scherer SW, Flanagan AM, Campbell PJ, Schiffman JD, Shago M, Alexandrov LB, Wunder JS, Andrulis IL, Malkin D, Behjati S, Shlien A. 2018. Rearrangement bursts generate canonical gene fusions in bone and soft tissue tumors. *Science* **361(6405)**:eaam8419 DOI 10.1126/science.aam8419.

Anisimova M, Gil M, Dufayard JF, Dessimoz C, Gascuel O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Systematic Biology* **60(5)**:685–699 DOI 10.1093/sysbio/syr041.

Bertolotti A, Melot T, Acker J, Vigneron M, Delattre O, Tora L. 1998. EWS, but not EWS-FLI-1, is associated with both TFIID and RNA polymerase II: interactions between two members of the TET family, EWS and hTAFII68, and subunits of TFIID and RNA polymerase II complexes. *Molecular and Cellular Biology* **18(3)**:1489–1497 DOI 10.1128/MCB.18.3.1489.

Bittrich S, Rose Y, Segura J, Lowe R, Westbrook JD, Duarte JM, Burley SK. 2021. RCSB protein data bank: improved annotation, search, and visualization of membrane protein structures archived in the PDB. *Bioinformatics* **38(5)**:1452–1454 DOI 10.1093/bioinformatics/btab813.

Chakrabortee S, Kayatekin C, Newby GA, Mendillo ML, Lancaster A, Lindquist S. 2016. Luminidependens (LD) is an Arabidopsis protein with prion behavior. *Proceedings of the National Academy of Sciences of the United States of America* **113(21)**:6065–6070 DOI 10.1073/pnas.1604478113.

Chang CK, Wu TH, Wu CY, Chiang MH, Toh EK, Hsu YC, Lin KF, Liao YH, Huang TH, Huang JJ. 2012. The N-terminus of TDP-43 promotes its oligomerization and enhances DNA binding affinity. *Biochemical and Biophysical Research Communications* **425(2)**:219–224 DOI 10.1016/j.bbrc.2012.07.071.

Chen C, Ding X, Akram N, Xue S, Luo SZ. 2019. Fused in sarcoma: properties, self-assembly and correlation with neurodegenerative diseases. *Molecules* **24(8)**:1622 DOI 10.3390/molecules24081622.

Chong PA, Vernon RM, Forman-Kay JD. 2018. RGG/RG motif regions in RNA binding and phase separation. *Journal of Molecular Biology* **430(23)**:4650–4665 DOI 10.1016/j.jmb.2018.06.014.

Conicella AE, Zerze GH, Mittal J, Fawzi NL. 2016. ALS mutations disrupt phase separation mediated by alpha-helical structure in the TDP-43 low-complexity C-terminal domain. *Structure* **24(9)**:1537–1549 DOI 10.1016/j.str.2016.07.007.

Couthouis J, Hart MP, Erion R, King OD, Diaz Z, Nakaya T, Ibrahim F, Kim HJ, Mojsilovic-Petrovic J, Panossian S, Kim CE, Frackelton EC, Solski JA, Williams KL, Clay-Falcone D, Elman L, McCluskey L, Greene R, Hakonarson H, Kalb RG, Lee VM, Trojanowski JQ, Nicholson GA, Blair IP, Bonini NM, Van Deerlin VM, Mourelatos Z, Shorter J, Gitler AD. 2012. Evaluating the role of the FUS/TLS-related gene EWSR1 in amyotrophic lateral sclerosis. *Human Molecular Genetics* **21(13)**:2899–2911 DOI 10.1093/hmg/dds116.

Cuevas-Velazquez CL, Dinneny JR. 2018. Organization out of disorder: liquid-liquid phase separation in plants. *Current Opinion in Plant Biology* **45(74)**:68–74 DOI 10.1016/j.pbi.2018.05.005.

Dasmeh P, Wagner A. 2021. Natural selection on the phase-separation properties of FUS during 160 my of mammalian evolution. *Molecular Biology and Evolution* **38(3)**:940–951 DOI 10.1093/molbev/msaa258.

Dosztanyi Z. 2018. Prediction of protein disorder based on IUPred. *Protein Science* **27(1)**:331–340 DOI 10.1002/pro.3334.

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. *Nucleic Acids Research* **47(D1)**:D427–D432 DOI 10.1093/nar/gky995.

Erdos G, Pajkos M, Dosztanyi Z. 2021. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Research* **49(W1)**:W297–W303 DOI 10.1093/nar/gkab408.

Fox NK, Brenner SE, Chandonia JM. 2014. SCOPe: structural classification of proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research* **42(D1)**:D304–D309 DOI 10.1093/nar/gkt1240.

Gu X, Chen Y, Wei Q, Cao B, Ou R, Yuan X, Hou Y, Zhang L, Liu H, Chen X, Shang HF. 2018. Mutation screening of the TIA1 gene in Chinese patients with amyotrophic lateral sclerosis/frontotemporal dementia. *Neurobiology of Aging* **68**:161.e1–161.e3 DOI 10.1016/j.neurobiolaging.2018.04.010.

Guindon S, Lethiec F, Duroux P, Gascuel O. 2005. PHYML Online–a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Research* **33(Web Server)**:W557–W559 DOI 10.1093/nar/gki352.

Harbi D, Harrison PM. 2014. Classifying prion and prion-like phenomena. *Prion* **8(2)**:161–165 DOI 10.4161/pri.27960.

Harrison PM. 2006. Exhaustive assignment of compositional bias reveals universally prevalent biased regions: analysis of functional associations in human and Drosophila. *BMC Bioinformatics* **7(1)**:441 DOI 10.1186/1471-2105-7-441.

Harrison PM. 2017. fLPS: fast discovery of compositional biases for the protein universe. *BMC Bioinformatics* **18(1)**:476 DOI 10.1186/s12859-017-1906-3.

Harrison PM. 2019. Evolutionary behaviour of bacterial prion-like proteins. *PLOS ONE* **14(3)**:e0213030 DOI 10.1371/journal.pone.0213030.

Harrison PM. 2021. fLPS 2.0: rapid annotation of compositionally-biased regions in biological sequences. *PeerJ* **9(16)**:e12363 DOI 10.7717/peerj.12363.

Iglesias V, Santos J, Santos-Suarez J, Pintado-Grima C, Ventura S. 2021. SGnn: a web server for the prediction of prion-like domains recruitment to stress granules upon heat stress. *Frontiers in Molecular Biosciences* **8**:718301 DOI 10.3389/fmolb.2021.718301.

Jackrel ME, Shorter J. 2014. Potentiated Hsp104 variants suppress toxicity of diverse neurodegenerative disease-linked proteins. *Disease Models & Mechanisms* **7**:1175–1184 DOI 10.1242/dmm.016113.

Kim HJ, Kim NC, Wang YD, Scarborough EA, Moore J, Diaz Z, MacLea KS, Freibaum B, Li S, Molliex A, Kanagaraj AP, Carter R, Boylan KB, Wojtas AM, Rademakers R, Pinkus JL, Greenberg SA, Trojanowski JQ, Traynor BJ, Smith BN, Topp S, Gkazi AS, Miller J, Shaw CE, Kottlors M, Kirschner J, Pestronk A, Li YR, Ford AF, Gitler AD, Benatar M, King OD, Kimonis VE, Ross ED, Weihl CC, Shorter J, Taylor JP. 2013. Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. *Nature* **495(7442)**:467–473 DOI 10.1038/nature11922.

Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simao FA, Zdobnov EM. 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research* **47(D1)**:D807–D811 DOI 10.1093/nar/gky1053.

Lancaster AK, Nutter-Upham A, Lindquist S, King OD. 2014. PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics* **30(17)**:2501–2502 DOI 10.1093/bioinformatics/btu310.

Law WJ, Cann KL, Hicks GG. 2006. TLS, EWS and TAF15: a model for transcriptional integration of gene expression. *Briefings in Functional Genomics and Proteomics* **5(1)**:8–14 DOI 10.1093/bfgp/ell015.

Letunic I, Bork P. 2018. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Research* **46(D1)**:D493–D496 DOI 10.1093/nar/gkx922.

Li HR, Chiang WC, Chou PC, Wang WJ, Huang JR. 2018. TAR DNA-binding protein 43 (TDP-43) liquid-liquid phase separation is mediated by just a few aromatic residues. *Journal of Biological Chemistry* **293(16)**:6090–6098 DOI 10.1074/jbc.AC117.001037.

Mackenzie IR, Bigio EH, Ince PG, Geser F, Neumann M, Cairns NJ, Kwong LK, Forman MS, Ravits J, Stewart H, Eisen A, McClusky L, Kretzschmar HA, Monoranu CM, Highley JR, Kirby J, Siddique T, Shaw PJ, Lee VM, Trojanowski JQ. 2007. Pathological TDP-43 distinguishes sporadic amyotrophic lateral sclerosis from amyotrophic lateral sclerosis with SOD1 mutations. *Annals of Neurology* **61**:427–434 DOI 10.1002/(ISSN)1531-8249.

Maruri-Lopez I, Figueroa NE, Hernandez-Sanchez IE, Chodasiewicz M. 2021. Plant stress granules: trends and beyond. *Frontiers in Plant Science* **12**:722643 DOI 10.3389/fpls.2021.722643.

Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang HY, El-Gebali S, Fraser MI, Gough J, Haft DR, Huang H, Letunic I, Lopez R, Luciani A, Madeira F, Marchler-Bauer A, Mi H, Natale DA, Necci M, Nuka G, Orengo C, Pandurangan AP, Paysan-Lafosse T, Pesseat S, Potter SC, Qureshi MA, Rawlings ND, Redaschi N, Richardson LJ, Rivoire C, Salazar GA, Sangrador-Vegas A, Sigrist CJA, Sillitoe I, Sutton GG, Thanki N, Thomas PD, Tosatto SCE, Yong SY, Finn RD. 2019. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research* **47(D1)**:D351–D360 DOI 10.1093/nar/gky1100.

Mitra J, Guerrero EN, Hegde PM, Liachko NF, Wang H, Vasquez V, Gao J, Pandey A, Taylor JP, Kraemer BC, Wu P, Boldogh I, Garruto RM, Mitra S, Rao KS, Hegde ML. 2019. Motor neuron disease-associated loss of nuclear TDP-43 is linked to DNA double-strand break repair defects. *Proceedings of the National Academy of Sciences of the United States of America* **116(10)**:4696–4705 DOI 10.1073/pnas.1818415116.

Molliex A, Temirov J, Lee J, Coughlin M, Kanagaraj AP, Kim HJ, Mittag T, Taylor JP. 2015. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell* **163(1)**:123–133 DOI 10.1016/j.cell.2015.09.015.

Monahan ZT, Rhoads SN, Yee DS, Shewmaker FP. 2018. Yeast models of prion-like proteins that cause amyotrophic lateral sclerosis reveal pathogenic mechanisms. *Frontiers in Molecular Neuroscience* **11**:453 DOI 10.3389/fnmol.2018.00453.

Naumann M, Pal A, Goswami A, Lojewski X, Japtok J, Vehlow A, Naujock M, Gunther R, Jin M, Stanslowsky N, Reinhardt P, Sterneckert J, Frickenhaus M, Pan-Montojo F, Storkebaum E, Poser I, Freischmidt A, Weishaupt JH, Holzmann K, Troost D, Ludolph AC, Boeckers TM, Liebau S, Petri S, Cordes N, Hyman AA, Wegner F, Grill SW, Weis J, Storch A, Hermann A. 2018. Impaired DNA damage response signaling by FUS-NLS mutations leads to

neurodegeneration and FUS aggregate formation. *Nature Communications* **9(1)**:335 DOI 10.1038/s41467-017-02299-1.

Nomura T, Watanabe S, Kaneko K, Yamanaka K, Nukina N, Furukawa Y. 2014. Intranuclear aggregation of mutant FUS/TLS as a molecular pathomechanism of amyotrophic lateral sclerosis. *Journal of Biological Chemistry* **289(2)**:1192–1202 DOI 10.1074/jbc.M113.516492.

Ohno T, Rao VN, Reddy ES. 1993. EWS/Fli-1 chimeric protein is a transcriptional activator. *Cancer Research* **53(24)**:5859–5863.

Ou SH, Wu F, Harrich D, Garcia-Martinez LF, Gaynor RB. 1995. Cloning and characterization of a novel cellular protein, TDP-43, that binds to human immunodeficiency virus type 1 TAR DNA sequence motifs. *Journal of Virology* **69(6)**:3584–3596 DOI 10.1128/jvi.69.6.3584-3596.1995.

Patel A, Lee HO, Jawerth L, Maharana S, Jahnel M, Hein MY, Stoynov S, Mahamid J, Saha S, Franzmann TM, Pozniakovski A, Poser I, Maghelli N, Royer LA, Weigert M, Myers EW, Grill S, Drechsel D, Hyman AA, Alberti S. 2015. A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell* **162(5)**:1066–1077 DOI 10.1016/j.cell.2015.07.047.

Picchiarelli G, Dupuis L. 2020. Role of RNA binding proteins with prion-like domains in muscle and neuromuscular diseases. *Cell Stress* **4(4)**:76–91 DOI 10.15698/cst2020.04.217.

Portz B, Lee BL, Shorter J. 2021. FUS and TDP-43 phases in health and disease. *Trends in Biochemical Sciences* **46(7)**:550–563 DOI 10.1016/j.tibs.2020.12.005.

Qin H, Lim LZ, Wei Y, Song J. 2014. TDP-43 N terminus encodes a novel ubiquitin-like fold and its unfolded form in equilibrium that can be shifted by binding to ssDNA. *Proceedings of the National Academy of Sciences of the United States of America* **111(52)**:18619–18624 DOI 10.1073/pnas.1413994112.

Ross ED, Baxa U, Wickner RB. 2004. Scrambled prion domains form prions and amyloid. *Molecular and Cellular Biology* **24(16)**:7206–7213 DOI 10.1128/MCB.24.16.7206-7213.2004.

Sephton CF, Cenik C, Kucukural A, Dammer EB, Cenik B, Han Y, Dewey CM, Roth FP, Herz J, Peng J, Moore MJ, Yu G. 2011. Identification of neuronal RNA targets of TDP-43-containing ribonucleoprotein complexes. *Journal of Biological Chemistry* **286(2)**:1204–1215 DOI 10.1074/jbc.M110.190884.

Sideri T, Yashiroda Y, Ellis DA, Rodriguez-Lopez M, Yoshida M, Tuite MF, Bahler J. 2017. The copper transport-associated protein Ctr4 can form prion-like epigenetic determinants in Schizosaccharomyces pombe. *Microbial Cell* **4(1)**:16–28 DOI 10.15698/mic2017.01.552.

Sievers F, Higgins DG. 2018. Clustal Omega for making accurate alignments of many protein sequences. *Protein Science* **27(1)**:135–145 DOI 10.1002/pro.3290.

Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N. 2010. PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Research* **38(suppl_1)**:D161–D166 DOI 10.1093/nar/gkp885.

Smethurst P, Sidle KC, Hardy J. 2015. Review: prion-like mechanisms of transactive response DNA binding protein of 43 kDa (TDP-43) in amyotrophic lateral sclerosis (ALS). *Neuropathology and Applied Neurobiology* **41(5)**:578–597 DOI 10.1111/nan.12206.

Sreedharan J, Blair IP, Tripathi VB, Hu X, Vance C, Rogelj B, Ackerley S, Durnall JC, Williams KL, Buratti E, Baralle F, de Belleroche J, Mitchell JD, Leigh PN, Al-Chalabi A, Miller CC, Nicholson G, Shaw CE. 2008. TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. *Science* **319(5870)**:1668–1672 DOI 10.1126/science.1154584.

Su TY, Harrison PM. 2019. Conservation of prion-like composition and sequence in prion-formers and prion-like proteins of saccharomyces cerevisiae. *Frontiers in Molecular Biosciences* **6**:54 DOI 10.3389/fmolb.2019.00054.

Su WC, Harrison PM. 2020. Deep conservation of prion-like composition in the eukaryotic prion-former Pub1/Tia1 family and its relatives. *PeerJ* **8(1)**:e9023 DOI 10.7717/peerj.9023.

Subramanian B, Gao S, Lercher MJ, Hu S, Chen WH. 2019. Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Research* **47(W1)**:W270–W275 DOI 10.1093/nar/gkz357.

Sun Z, Diaz Z, Fang X, Hart MP, Chesi A, Shorter J, Gitler AD. 2011. Molecular determinants and genetic modifiers of aggregation and toxicity for the ALS disease protein FUS/TLS. *PLOS Biology* **9(4)**:e1000614 DOI 10.1371/journal.pbio.1000614.

Wang J, Choi JM, Holehouse AS, Lee HO, Zhang X, Jahnel M, Maharana S, Lemaitre R, Pozniakovsky A, Drechsel D, Poser I, Pappu RV, Alberti S, Hyman AA. 2018. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell* **174(3)**:688–699.e16 DOI 10.1016/j.cell.2018.06.006.

Wang Y, Harrison PM. 2021. Homopeptide and homocodon levels across fungi are coupled to GC/AT-bias and intrinsic disorder, with unique behaviours for some amino acids. *Scientific Reports* **11(1)**:10025 DOI 10.1038/s41598-021-89650-1.

Wang WY, Pan L, Su SC, Quinn EJ, Sasaki M, Jimenez JC, Mackenzie IR, Huang EJ, Tsai LH. 2013. Interaction of FUS and HDAC1 regulates DNA damage response and repair in neurons. *Nature Neuroscience* **16(10)**:1383–1391 DOI 10.1038/nn.3514.

Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of Molecular Biology* **337(3)**:635–645 DOI 10.1016/j.jmb.2004.02.002.

Yuan AH, Hochschild A. 2017. A bacterial global regulator forms a prion. *Science* **355(6321)**:198–201 DOI 10.1126/science.aai7776.

Zajkowski T, Lee MD, Mondal SS, Carbajal A, Dec R, Brennock PD, Piast RW, Snyder JE, Bense NB, Dzwolak W, Jarosz DF, Rothschild LJ. 2021. The hunt for ancient prions: archaeal prion-like domains form amyloid-based epigenetic elements. *Molecular Biology and Evolution* **38(5)**:2088–2103 DOI 10.1093/molbev/msab010.

Zhang K, Liu Q, Shen D, Tai H, Fu H, Liu S, Wang Z, Shi J, Ding Q, Li X, Liu M, Cui L, Zhang X. 2018. Genetic analysis of TIA1 gene in Chinese patients with amyotrophic lateral sclerosis. *Neurobiology of Aging* **67**:201.e9–201.e10 DOI 10.1016/j.neurobiolaging.2018.03.020.

Zhao M, Kim JR, van Bruggen R, Park J. 2018. RNA-binding proteins in amyotrophic lateral sclerosis. *Molecules and Cells* **41**:818–829 DOI 10.14348/molcells.2018.0243.