The mutational landscape of a prion-like domain

Benedetta Bolognesi\textsuperscript{1,2,6}, Andre J. Faure\textsuperscript{\textcircled{6}}, Mireia Seuma\textsuperscript{1,2}, Jörn M. Schmiedel\textsuperscript{1}, Gian Gaetano Tartaglia\textsuperscript{1,3,4,5} & Ben Lehner\textsuperscript{1,3,4}

Insoluble protein aggregates are the hallmarks of many neurodegenerative diseases. For example, aggregates of TDP-43 occur in nearly all cases of amyotrophic lateral sclerosis (ALS). However, whether aggregates cause cellular toxicity is still not clear, even in simpler cellular systems. We reasoned that deep mutagenesis might be a powerful approach to disentangle the relationship between aggregation and toxicity. We generated >50,000 mutations in the prion-like domain (PRD) of TDP-43 and quantified their toxicity in yeast cells. Surprisingly, mutations that increase hydrophobicity and aggregation strongly decrease toxicity. In contrast, toxic variants promote the formation of dynamic liquid-like condensates. Mutations have their strongest effects in a hotspot that genetic interactions reveal to be structured in vivo, illustrating how mutagenesis can probe the in vivo structures of unstructured proteins. Our results show that aggregation of TDP-43 is not harmful but protects cells, most likely by titrating the protein away from a toxic liquid-like phase.

\textsuperscript{1}Center for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Doctor Aiguader 88, 08003 Barcelona, Spain. \textsuperscript{2}Institute of Bioengineering of Catalonia (IBEC), The Barcelona Institute of Science and Technology, Barcelona, Spain. \textsuperscript{3}Universitat Pompeu Fabra (UPF), Barcelona, Spain. \textsuperscript{4}Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain. \textsuperscript{5}Department of Biology ‘Charles Darwin’, Sapienza University of Rome, P.le A. Moro 5, Rome 00185, Italy. \textsuperscript{6}These authors contributed equally: Benedetta Bolognesi, Andre J. Faure. Correspondence and requests for materials should be addressed to B.B. (email: bbolognesi@ibecbarcelona.eu) or to B.L. (email: ben.lehner@crg.eu)
The conversion of specific proteins into insoluble aggregates is a hallmark of many neurodegenerative disorders, including Alzheimer’s, Parkinson’s, Huntington’s, and Amyotrophic Lateral Sclerosis (ALS) with dominantly inherited mutations in aggregate-forming proteins causing rare familial forms of these diseases. However, both in humans and in animal models, there is often only a weak association between the presence of aggregates and disease progression. Indeed, multiple therapeutic approaches that reduce the formation of aggregates have failed at different stages of development. On the other hand, there is increasing evidence that alternative protein mechanisms, including the formation of insoluble aggregates, the latter are still widely assumed to be pathogenic in many neurodegenerative diseases.

For many proteins, aggregation depends critically on intrinsically disordered regions with a low sequence complexity resembling that of infectious yeast prions. These prion-like domains (PRDs) are also enriched in proteins that can form liquid-like cellular condensates through liquid-demixing. This is a concentration-dependent process through which proteins can separate into two coexisting liquid phases and it has been extensively characterized both in vitro and in the cytoplasm. In several proteins PRDs are necessary and sufficient for liquid-liquid demixing. At least in vitro, insoluble aggregates can nucleate from more liquid phases, leading to the suggestion that liquid de-mixed states can mature into pathological aggregates.

Disordered regions and low-complexity sequences are also enriched in dosage-sensitive proteins that are toxic when their concentration is increased. At least for one model protein that has been tested, however, it is the formation of a concentration-dependent liquid-like phase—not aggregation—that causes cellular toxicity. Similarly, the toxicity of two mutant forms of the prion Sup35 could be explained only on the basis of their ability to populate a non-aggregate, liquid-like state.

Cytoplasmic aggregates of the TAR DNA-binding protein 43 (TDP-43) are a hallmark of ALS, present in 97% of post-mortem samples. TDP-43 aggregates are also present at autopsy in nearly all cases of frontotemporal dementia (FTD) that lack tau-containing inclusions (about half of all cases of FTD which is the second most common dementia). TDP-43 aggregates also represent a hallmark of inclusion body myopathy, and a secondary pathology in Alzheimer’s, Parkinson’s, and Huntington’s diseases. However, TDP-43 aggregates are also observed—albeit at low frequency—in control samples and, in vitro, TDP-43 can form both amyloid aggregates and liquid condensates. Mutations in TDP-43 cause ~5% of familial ALS (fALS) cases, with these mutations reported to interfere with nuclear-cytoplasmic transport, RNA processing, splicing, and protein translation. However, despite extensive investigation, the molecular form of the protein that causes cellular toxicity is still unknown.

We reasoned that systematic ‘deep’ mutagenesis could be an unbiased approach to identify and investigate the toxic species of proteins. A map of which amino acid (AA) changes increase or decrease the toxicity of a protein to a cell should, if sufficiently comprehensive, clarify both the properties of the protein and its in vivo conformational states associated with toxicity. The effects of a small number of mutations on TDP-43 toxicity or aggregation have been previously reported. However, on the basis of a handful of mutations, the relationship between aggregation and toxicity is far from clear.

Here we show by quantifying the effects of >50,000 mutations in the PRD of TDP-43 that increasing hydrophobicity and aggregation strongly reduce the toxicity of this protein in yeast. Moreover, mutations that increase the toxicity of TDP-43 actually promote the formation of dynamic liquid-like cytoplasmic condensates. Mutations have their strongest effects in a central ‘hotspot’ region of the PRD TDP-43. The patterns of genetic interactions in double mutants in this region reveal that this ‘unstructured’ region is actually structured in vivo. Our results illustrate how deep mutagenesis can be used to probe the sequence-function relationships and the in vivo structures of ‘disordered’ proteins. We propose that aggregation of TDP-43 is not harmful but actually protects cells, most likely by titrating protein from a toxic liquid-like phase.

Results

Deep mutagenesis of the TDP-43 prion-like domain. We used error-prone oligonucleotide synthesis to comprehensively mutate the PRD of TDP-43. We introduced the library into Saccharomyces cerevisiae cells, induced expression and used deep sequencing before and after induction to quantify the relative effects of each variant on growth in three biological replicates (Fig. 1a). After quality control and filtering (Supplementary Fig. 1a and c), the dataset quantifies the relative toxicity of 1,266 single and 56,730 double amino acid (AA) changes in the PRD with high reproducibility (Fig. 1b, Supplementary Fig. 1d and e). The toxicity scores also correlate very well with the toxicity of the same variants re-tested in the absence of competition (Fig. 1c).

The toxicity of both single and double mutants has a tri-modal distribution (Fig. 1d, Supplementary Fig. 2a and c), with 18,023 variants more toxic and 16,152 variants less toxic than wild-type (WT) TDP-43 (t-test false discovery rate, FDR = 0.05). The dataset therefore allows us to investigate how mutations both increase and decrease toxicity. Very interestingly, ALS TDP-43 mutations increase toxicity, with a strong bias towards moderate effects (t-test, p-value = 0.005) (Fig. 1d, Supplementary Fig. 2d).

Mutation effects are largest in a central hotspot of the PRD. Plotting the mean toxicity of all mutations at each position in the sequence reveals a 31 AA hotspot (312–342) where the effects of mutations are strongest (Fig. 1e). The variance in toxicity per position is also the highest within this hotspot, with mutations both strongly increasing and decreasing toxicity (Fig. 1e). A heatmap of the toxicity of all of the single mutations also clearly reveals this hotspot, with most mutations of strong positive or negative effect falling within this 31 AA window (Fig. 1f). Equally strikingly, mutations to the same AA but in different positions within the hotspot often have very similar effects (Fig. 1f). In particular, mutations to charged and polar residues increase toxicity throughout the hotspot and mutations to hydrophobic AAs decrease toxicity (Fig. 1f).

Hydrophobicity and aggregation potential predict toxicity. To more systematically identify features associated with changes in toxicity we made use of all 53,468 variants carrying one or two AA substitutions (excluding STOP codon variants). We used principal components analysis (PCA) to reduce the redundancy in a list of over 350 AA physicochemical properties (Supplementary Fig. 3) and linear regression to quantify how well changes in these physicochemical properties predict changes in the toxicity of TDP-43. A principal component very strongly related to hydrophobicity is the most predictive feature of toxicity, explaining 66% of the variance in toxicity of all 8,040 mutants within the 312–342 hotspot and 51% of the variance in toxicity of all genotypes (Fig. 2a). With the same approach, we tested the performance of established predictors of protein aggregation, intrinsic disorder and other properties. None of them are as predictive as hydrophobicity (Fig. 2b). Importantly, after controlling for hydrophobicity, additional features such as...
charge and aromaticity do not predict toxicity (Fig. 2d, e, Supplementary Fig. 4a) with aggregation potential accounting for an additional 4% of variance in the hotspot (Fig. 2f, g). That increased hydrophobicity and aggregation potential are strongly associated with reduced toxicity across >50,000 genotypes was unexpected given previous work that reported an increased number of intracellular aggregates for a set of TDP-43 variants toxic to yeast and the widely-held view that aggregation is harmful to cells. We therefore further investigated the effects of mutants that alter the hydrophobicity and toxicity of TDP-43.

| Relative toxicity Estimates | Mutant effect hotspot | Heatmap showing single mutant toxicity estimates. The horizontal dashed line indicates the mean absolute toxicity of all single mutants. |
|---------------------------|----------------------|----------------------------------------------------------------------------------------------------------------------------------|
| Relative toxicity w.r.t. WT normalized by doubling time (replicate1) | Relative toxicity w.r.t. WT normalized by doubling time (replicate2) | Density |
| -0.4 | 0.0 | 0.4 |
| -0.2 | 0.0 | 0.2 |
| 0.0 | 0.0 | 0.0 |
| 0.2 | 0.0 | 0.2 |
| 0.4 | 0.0 | 0.4 |

Fig. 1 Deep mutational scanning (DMS) of the prion-like domain (PRD) of TDP-43. a Domain structure of TDP-43 and DMS experimental protocol: For each library, three independent selection experiments were performed. In each experiment one input culture was split into two cultures for selection upon induction of TDP-43 expression (6 outputs total). Relative toxicity of variants was calculated from changes of output to input frequencies relative to WT. b Correlation of toxicity estimates between replicates 1 and 2 for single and double amino acid (AA) mutants shown separately for each library (290–332; 332–373). The Pearson correlation coefficients (R) are indicated. Toxicity correlations between all replicates are shown in Supplementary Fig. 1d, e. c Comparison of toxicity from pooled selections and individually measured growth rates for selected variants. Vertical and horizontal error bars indicate 95% confidence intervals of mean growth rates and toxicity estimates respectively. Linear fits of the data are shown separately for each library and Pearson correlation (R) after pooling data from both libraries is indicated. d Toxicity distribution of single and double mutants, shown separately for each library (colour key as in panel (c)). WT variant has toxicity of zero, mean toxicity of variants with single STOP codon mutation is indicated by dashed vertical line. The red boxplot depicts the distribution of toxicity estimates for all human disease mutations (including sporadic and familial ALS mutations). Outliers are not depicted but are reported in Supplementary Fig. 2d, e. e Absolute toxicity of single mutants stratified by position. Error bars indicate 95% confidence intervals of mean per-position toxicity estimates. A local polynomial regression (loess) over toxicity estimates of all single mutants is shown. The vertical dashed line indicates the boundary between the two DMS libraries. The horizontal dashed line indicates the mean absolute toxicity of all single mutants. The mutant effect “hotspot” (mean per-position toxicity > mean toxicity) is highlighted in grey. f Heatmap showing single mutant toxicity estimates. The vertical axis indicates the identity of the substituted (mutant) AA. Heatmap cells of variants not present in the library are denoted by "."
Changes in hydrophobicity are highly predictive of TDP-43 cellular toxicity. 

**a** Percentage variance of toxicity explained by linear regression models predicting single and double mutant variant toxicity from changes in AA properties upon mutation (PCs, principal components of a collection of AA physico-chemical properties). Different regression models were built for different subsets of the data. Simple linear regression models for all variants (blue) or only variants inside (red) or outside (yellow) the hotspot region. And a regression model using all variants and including a binary location variable (inside/outside hotspot) as well as an interaction term between binary location variable and the indicated AA property feature (green). 

**b** Percentage variance of toxicity explained by linear regression models predicting variant toxicity using scores from aggregation/structure algorithms (see Methods). Colour key shown in panel (a). See also Supplementary Fig. 4. 

**c** Toxicity of variants with single or double mutations within the hotspot region as a function of hydrophobicity changes (PC1) induced by mutation. The Pearson correlation (R) before binning is indicated. See also Supplementary Fig. 9a. 

**d** Toxicity distributions of single and double mutants stratified by the change in the number of aromatic (H,F,W,Y,V) or charged residues (R,D,E,K) relative to the WT sequence. Horizontal axis as in panel (e). 

**e** Distribution of residual toxicity after controlling for the effect of hydrophobicity and location on toxicity (green regression model in panel a) stratified by the number of aromatic (H,F,W,Y,V) or charged (R,D,E,K) AAs. 

**f** Single and double mutant variant toxicity as a function of changes in aggregation propensity (Zyggregator). Only variants occurring within the toxicity hotspot are depicted. The Pearson correlation (R) before binning is indicated. 

**g** Toxicity as a function of aggregation propensity after controlling for hydrophobicity (red regression model in panel a). Only variants occurring within the toxicity hotspot are depicted. The Pearson correlation (R) before binning is indicated. See also Supplementary Fig. 9b.
**Two classes of cytoplasmic TDP-43 foci.** WT TDP-43 localizes to both the nucleus and to the cytoplasm of yeast cells \(^{54,55}\) (Fig. 3a). In the nucleus, TDP-43 is diffuse, but in the cytoplasm it forms puncta, consistent with previous observations \(^{41,57}\). We observe that cytoplasmic WT TDP-43 forms two types of assemblies: small foci in the nuclear periphery and larger foci detached from the nucleus (Fig. 3a, c). We find that mutations that decrease TDP-43 hydrophobicity and increase TDP-43 toxicity increase the number of the small foci at the nuclear periphery and reduce the number of large distal foci (Fig. 3b, c, f, g, h).
Fig. 3 Mutations leading to formation of solid-like aggregates rescue toxicity. a Representative fluorescence microscopy images of yeast cells expressing indicated YFP-tagged TDP-43 variants (W334K TDP-43 = toxic, A328V TDP-43 = non-toxic). H4-mCherry marks nuclei (red). Contrast was enhanced equally for the green and red channels in all images. b Percentage of cells with cytoplasmic foci (Cells scored: n[toxic] = 219, n[WT] = 30, n[non-toxic] = 213). Fisher’s Exact test. c Percentage of cells with cytoplasmic foci with size over 5 pixels automatically detected by CellProfiler. Fisher’s Exact test. (Cells scored: n[toxic] = 167, n[WT] = 23, n[non-toxic] = 167). d Percentage of cells with foci at the nuclear periphery (Cells scored: n[toxic] = 219, n[WT] = 30, n[non-toxic] = 213). Fisher’s exact test. e Distance of foci from nucleus center for toxic (red), non-toxic (blue), and WT (black) TDP-43. Boxplots represent median values, interquartile ranges and Tukey whiskers with individual data points superimposed. Mann-Whitney test. f Average fluorescence intensity of foci localized closer (<15 pixels, n=147) or further (>15 pixels, n=138) from the nucleus. Boxplots represent median values, interquartile ranges and Tukey whiskers with individual data points superimposed. Mann-Whitney test. g Representative individual fluorescence recovery traces for variants reported in panel (e). Lines are the result of a single exponential fitting. h Mobile Fraction as calculated by fitting FRAP traces for toxic (red), non-toxic (blue) and WT (black) TDP-43. Each point results from fitting an individual trace. One-way ANOVA with Tukey’s multiple comparisons test. Images were taken on cells growing from at least 3 independent starting colonies. *P<0.05, **P<0.01, ***P<0.001, ****P<0.0001. Scale bar = 5 μM. Source data are provided as a Source Data file.

Supplementary Fig. 5a). TDP-43 mutations reported in ALS (Supplementary Fig. 2e) also increase the number of foci at the nuclear periphery compared to WT TDP-43 (Supplementary Fig. 6a, b). In contrast, mutations that increase hydrophobicity and reduce toxicity reduce the number of small nucleus-associated foci and increase the number of large distal foci (Fig. 3b, c, f, Supplementary Fig. 5a).

Toxic mutations promote dynamic liquid-like condensates. We used fluorescence recovery after photobleaching (FRAP) to characterize the dynamics of TDP-43 variants in the different foci. The large cytoplasmic foci formed by non-toxic variants show little exchange of TDP-43 molecules with the soluble cytoplasmic pool. In contrast, the small foci localized at the nuclear periphery can exchange more protein with the cytoplasm, consistent with a more liquid-like state (Fig. 3d, e). Such differences in dynamics have been described also for distinct types of misfolded protein compartments58. Both types of compartments co-localize with the yeast chaperone Hsp104 (Supplementary Fig. 7a). The large immobile TDP-43 foci are also brighter than the small dynamic ones (Fig. 3g), similar to what has been observed for Huntingtin variants that partition between immobile bright assemblies and liquid-like dimmer ones59. The non-toxic TDP-43 variants also have a higher protein concentration quantified by Western blotting (Supplementary Fig. 5b).

Taken together, these results suggest that mutations that increase the hydrophobicity of TDP-43 result in a re-localization of the protein away from small and dynamic, liquid-like foci at the nuclear periphery to large and more solid aggregates in the cytoplasm. A reduction in hydrophobicity has the opposite effect.

Genetic interactions reveal the hotspot structure in vivo. The hotspot region of the TDP-43 PRD (AA 312-342) is a conserved region35,36, with hydrophobicity more similar to the globular domains of TDP-43 than to the surrounding hydrophilic disordered regions (Fig. 4b). The hotspot is contained within a region (311-360) that was previously shown to be sufficient for both in vitro aggregation and the formation of cytoplasmic foci35. Fragments from within this region have previously been shown to have the potential to form different types of secondary structures in vitro. More specifically, nuclear magnetic resonance (NMR) spectroscopy of the PRD revealed that residues 321–342 can adopt an α-helical structure in certain conditions35,36,47 and four different 6-11 AA peptides from the region could form cross-β amyloid or amyloid-like fibrils whose structures were determined by X-ray crystallography52. However, it is unknown whether any of these structures exist in vivo for full-length TDP-43.

We have shown recently that the pattern of genetic (epistatic) interactions between mutations in a protein can report on the secondary structure of that molecule when it is performing the function that is being selected for51,60. In particular, when a sequence forms an α-helix, the side chains of residues separated by 3–4 AA are close in space and similarly oriented so that mutations in these AA interact similarly with mutations in the rest of the protein. In contrast, in a β-strand, the side chains of residues separated by 2 AA are close and similarly oriented and so make similar genetic interactions with other mutations (Fig. 4a)61.

We used the 52,272 double mutants (excluding STOP codon variants) in our dataset to identify pairs of mutations that genetically interact. We first identified pairs of mutations that had unexpectedly high or low toxicity (<5th and >95th percentile of the expected toxicity distribution, negative and positive epistasis for growth rate, respectively). We then quantified the similarity of epistasis enrichment profiles between pairs of positions and compared these patterns to those expected for α-helices and β-strands, scoring significance by randomization51 (Fig. 4a).

This revealed that the patterns of epistasis in our dataset are consistent with two secondary structure elements forming inside the PRD in vivo: a β-strand at residues 311–316 and an α-helix at residues 324–331 (Fig. 4c). The β-strand identified by the epistasis analysis coincides with one of the peptides in the TDP-43 PRD that, in vitro, can form cross-β structures52 typical of protein aggregates (Fig. 4d). The crystals of this specific peptide consist of a non-conventional β-strand termed a low-complexity aromatic-rich kinked segment (LARKS)62. In this in vitro structure, Phe 313 and Phe 316 face the same side of the sheet, whereas in a canonical sheet the side chains of odd and even residues face opposite sides. Strikingly, this non-canonical contact between Phe 313 and Phe 316 is also identified by the in vivo epistasis analysis, with a similarity in interaction profile ranking amongst the top two residue pairs in this region. In addition, the contact between Phe 316 and Ala 315, which again is compatible with a LARKS but not with a canonical β-strand has the highest predicted contact score among neighbouring residues (Fig. 4d).

The predicted contact map built on the basis of in vivo epistatic interactions strikingly matches the Protein Data Bank (PDB) structure for LARKS 312–317 (Fig. 4d, Supplementary Fig. 8).

On the other hand, the genetic interactions of mutations in the 324–330 region match those expected for an α-helix (Fig. 4e). This region is part of the portion (321–342) of TDP-43 that can transiently and cooperatively fold into an α-helix in vitro36,47,63. This helix is stabilized by inter-molecular contacts and its self-interaction was proposed to seed liquid-demixing in vitro. Amyloid fibrils can grow from the liquid de-mixed state and circular dichroism spectroscopy revealed that the helix can transition to a β-sheet over time, compatible with the process of aggregation55,63. On the basis of epistasis, the top scoring predicted contacts in this region are between residues separated...
Correlated patterns of epistasis predict secondary structural elements within the PRD of TDP-43. a Schematic representation of the computational strategy to identify in vivo secondary structures. Double mutant variants are classified as epistatic if they are more (95th percentile) or less (5th percentile) toxic than other variants with similar single mutant toxicities (top). A pair-wise interaction (PWI) matrix of epistasis correlation scores is then constructed by quantifying the similarity of a pair of positions’ interactions with all other mutated positions in the protein. The epistasis correlation scores along the diagonal of the PWI matrix are then tested for agreement with the stereotypical periodicity of α-helix and β-strand, using two-dimensional kernels (bottom), to calculate the likelihood of adjacent positions forming secondary structures. b Local polynomial regression (loess) of hydrophobicity (PCI) of the WT TDP-43 sequence with 95% confidence interval. For reference, smoothed toxicity estimates in the mutated positions within the PRD are shown. The Pearson correlation coefficient (R) between hydrophobicity and mean toxicity effects of single mutants at each position before smoothing is indicated. For reference, smoothed toxicity estimates in the mutated positions within the PRD are shown. The Pearson correlation coefficient (R) between hydrophobicity and mean toxicity effects of single mutants at each position before smoothing is indicated. c Secondary structure prediction matrix along the diagonal of the PWI matrix by 3–4 AA such as Ala 324 and Ala 328, or Ala 325 and Ala 328, consistent with interactions between side chains of an α-helix (Fig. 4e).

The pattern of in vivo epistatic interactions between mutations in TDP-43 therefore is compatible with a model in which two of the secondary structures that have previously been observed
in vitro for fragments of TDP-43 actually form in vivo in the full-length protein.

Discussion
Specific protein aggregates have long been recognized as the hallmarks of many neurodegenerative diseases\(^1\-6,32,64\). However, whether these aggregates are the cause of these diseases, non-pathological by-products, or a protective mechanism is still very unclear and hotly debated\(^13\-16\). Indeed, although it is often assumed to be the case, it is not even clear whether aggregates are the cause of toxicity when aggregating proteins are expressed in simpler cellular systems\(^34,55\). We reasoned that deep mutagenesis might be an effective approach to resolve this question.

In this study, we have tested this approach using the ALS protein TDP-43 that both aggregates and causes toxicity in the model eukaryote, S. cerevisiae. Quantifying the effects of >50,000 mutants of TDP-43 revealed unequivocally that increasing the hydrophobicity and aggregation of TDP-43 strongly reduces the toxicity of this protein in yeast cells. Consistently, mutations that reduce hydrophobicity and the aggregation potential of TDP-43 increase the toxicity of the protein. Although they reduced the formation of large, solid aggregates, mutations that increase toxicity promote the formation of alternative foci—dynamic, liquid-like TDP-43 condensates clustered at the nuclear periphery. We propose therefore that aggregation reduces the toxicity of TDP-43 to yeast cells because it titrates TDP-43 away from this toxic liquid-like phase (Fig. 5a).

That TDP-43 aggregates are protective rather than toxic is consistent with previous work in multiple systems, including the rescue of toxicity by the accumulation of RNA lariats that sequester TDP-43 into large aggregates\(^65\). Moreover, in mammalian cells, liquid de-mixed TDP-43 was recently shown to recruit the nuclear pore component Nup62 and the importin-\(\alpha\) transporter, resulting in nuclear transport impairment and toxicity\(^54\). Thus, although it still remains to be established whether aggregation of TDP-43 is also protective in mammalian cells and neurons, it seems likely that this will be the case. The observation that all recurrent fALS mutations increase the toxicity of TDP-43 in yeast and by a similar magnitude (Supplementary Fig. 2d) is very striking and suggests that the yeast system may indeed capture molecular mechanisms relevant to the human disease. Indeed, given the late age of onset of ALS, it is particularly interesting that the fALS mutations are all moderate effect mutations when expressed in yeast, as it may be the case that the more toxic variants of TDP-43 are embryonic lethal in humans.

More generally, our results demonstrate that deep mutagenesis is a powerful approach for determining the sequence-function relationships of intrinsically disordered proteins, including probing their in vivo structures. Mutations had their strongest effects within a central hotspot region of the TDP-43 PRD. Our recently developed approach\(^51\) that uses the patterns of genetic interactions in double mutants to report on structural contacts reveals that this ‘unstructured’ hotspot region is very likely to be structured in vivo with the formation of these secondary structures altering the toxicity of the protein. Indeed, secondary structure elements within this region have been shown to be important for the phase separation and aggregation of fragments of TDP-43 in vitro\(^35,36,52\). A parsimonious model based on previous in vitro work\(^35,36,47\) is that the helix forms first in the pathway of aggregation towards a \(\beta\)-rich species (Fig. 5b). Consistent with this, destabilizing mutations, such as any substitution of Phe 313 and Phe 316 in the LARKS, or the introduction of proline into the 324–330 helix, increase toxicity (Fig. 1f).

The conformations of ‘unstructured’ proteins are notoriously difficult to study and the interactions between mutations in double mutants provide a general method to probe the in vivo structures of these proteins whenever a selection assay is available. We envisage that this approach can be adopted to study the functions, toxicity, and in vivo structures of other intrinsically disordered proteins, including the many other proteins implicated in neurodegenerative diseases.

Our conclusions derived from deep mutagenesis of TDP-43 are also consistent with observations for other genes, such as the reduced toxicity of SOD-1 variants that increase aggregation\(^16,66\) and the increased survival of neurons containing Huntingtin inclusion bodies\(^67\). They are also consistent with increasing evidence that insoluble aggregates are not pathogenic in multiple other neurodegenerative diseases\(^64,68,69\), and with the clinical failure of therapeutic approaches that reduce the occurrence of aggregates\(^10,12,70\-72\).

Indeed, if insoluble aggregates generally titrate proteins away from alternative toxic phases, interactions and functions, then promoting rather than alleviating aggregation might be the more appropriate therapeutic goal in neurodegenerative diseases.

Methods
Yeast strains and plasmids. Saccharomyces cerevisiae S288C BY4741 (MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0) was used in all experiments. Plasmid pRS416 containing TDP-43 or TDP-43-YFP under control of the Gal1 promoter was purchased from Addgene\(^65\). Mutagenesis for the characterization of TDP-43 variants was performed through PCR linearization with specifically-designed primers (Supplementary Data 1, primers: BB_1 to BB_6). The resulting products were then either treated with DpnI or purified from a 1% agarose gel with a QIAquick Gel Extraction Kit (Qiagen) and transformed into E.coli DH5α competent cells.
paired-end Illumina sequencing at the CRG Genomics Unit. All replicates were pooled together in an equimolar ratio. Finally, the pooled DNA was amplified by PCR (Q5 High-Fidelity DNA Polymerase, NEB) for 15 cycles, purified using an E-gel electrophoresis system (Agarose 2%) followed by column purification with a MinElute PCR Purification Kit (Qiagen). In order to introduce the doped sequence in the full-length TDP-43 sequence, the purified oligonucleotide was cloned into 100 mg of linearized pRS416 Gal TDP-43 by a Gibson approach (Supplementary Data 1, primers BB_7 to BB_10). The product was then transformed into 10-beta Electrocompetent E. coli (NEB), by electroporation in a Bio-Rad GenePulser machine (2.0 kV, 200 μL 25 μF). Cells were recovered in SOC medium (NEB) for 30 min and plated on LB with ampicillin. A total of ~2 × 10^6 transformants were estimated. The plasmid library was purified with a Kit (THERMOScientific®). Yeast cells were transformed with TDP-43 doped plasmid in 4 independent biological replicates for each library. One single colony was grown overnight in 30 mL YPDA medium at 30 °C for each replica. Cells were diluted to 0.5 optical density at a wavelength of 600 nm (OD600) in 175 mL of YPDA and incubated for 4 h at 30 °C. Cells were then harvested, washed, re-suspended in 8.575 mL SORB (100 mM LiOAc, 10 mM Tris pH 8.0, 1 mM EDTA, 1 mM sorbitol) and incubated for 30 min at room temperature. For the transformation, 10 mg per mL of salmon sperm DNA and 3.5 μg TDP-43 plasmid library were used. Cells were mixed to 100 mM LiOAc, 10 mM Tris-HCl pH 8.0, 1 mM EDTA/NaOH pH 8.0 and 40% PEG 3350. Heat-shock was performed for 20 min at 42 °C. YPD with 0.5 M sorbitol was used to recover the cells, incubating them for 1 h at 30 °C. After recovery, cells were resuspended in SC-URA 2% raffinose medium, while an aliquote was plated to calculate transformation efficiency.

After ~50 h of growth, cells were diluted in SC-URA 2% raffinose medium and grown for 4.5 generations. At this stage, 400 mL of each replica were harvested, washed, split into two tubes and frozen at −20 °C for later extraction of input DNA. To induce plasmid expression, for each replicate two cultures were diluted in 200 μL of galactose non-inducing medium and then diluted to 0.2 OD600 in galactose medium to induce protein expression for 8 h. The cells were immobilized to an 8-well cover slide by Concanavalin-A-mediated cell adhesion. Cells were then imaged under a Confocal TCS SP8 microscope (Leica) where bleaching was achieved with 488 Laser Power at 70% for three frames (1.3 s per frame) while fluorescence recovery was recorded for 50 frames. The curves were then fitted to a single exponential, following normalization, with the EasyFrame package.

Protein extraction and western blotting. Single yeast colonies were grown overnight in non-inducing medium and then diluted to 0.2 OD600 in galactose medium to induce protein expression for ~8 h. At this stage, 6 × 10^6 cells were collected and re-suspended in 200 μL EDTOH and 2.5 μL PMSF. Samples were vortexed with glass beads for 15 min at 4 °C and frozen overnight at −80 °C. The samples were dried in a speed vacuum and re-suspended in 200 μL of solubilizing buffer (20 mM Tris pH 6.8, 2% SDS). After boiling for 5 min, the lysate fraction was run on a NuPAGE 4–12% Bis-Tris gels (Novex) and transferred to PVDF membranes in an iBlot (Invitrogen). Membranes were blocked with 5% milk powder in TBS-T and incubated overnight at 4 °C with antibodies: anti-GFP mouse antibody (Santa Cruz sc-9996) and anti-PGKD1 mouse antibody (Novex 459250) diluted 1:1000 and 1:5000 in 2.5% powdered milk respectively. Secondary antibody anti-proteinG was incubated for 1 h at room temperature. Proteins were detected with an enhanced chemi-luminescence system (Millipore Luminata) and visualized using an Amersham Imager 600 (GE Healthcare).

Sequencing data pre-processing. FastQ files from paired-end sequencing of replicate deep mutational scanning (DMS) libraries before (’input’) and after selection (’output’) were processed using a custom pipeline (https://github.com/lehner-lab/tardbpdms_cellprofiler_scripts). First, 5’ constant regions were trimmed, but read pairs were discarded if 5’ constant regions contained more than 20% mismatches to the reference sequence. Read pairs were aligned (reads that did not match the expected 126 bp length were discarded) and Phred base quality scores of aligned positions were calculated using USEARCH. Reads that contained base calls with Phred scores below 30 (290–331 DMS library) or below 25 (332–373 DMS library) were discarded. Approximately five and seven million reads passed these filtering criteria in each sample corresponding to the 290–331 and 332–373 libraries respectively. Finally, unique variants were counted and merged into a single table of variant counts (aggregated across technical output replicates) per DMS library. One out of four input replicates (and three from each DMS library) was discarded due to considerably lower correlations with the other replicates (Supplementary Fig. 1a, b).
**Variant toxicity and error estimates.** All analyses of toxicity were performed on variants with a maximum of two AA mutations, but no synonymous mutations in other codons. Error estimates for per AA sequence indices were calculated by first replicating and merging paired (i.e. BETareADIR, DISEMBL, IUPRED2A, etc.) results on the output from ZipperDB, Zyggregator. The entire PRD AA sequence was supplied to AGADIR and all unique six-mers to ZipperDB. For the remainder, the full-length AA sequence was used.

Variants inside the hotspot were defined as those with mutant residue positions in the range of 312–342. Change in absolute charge (regardless of sign) is shown in Fig. 2d, e, because this feature is more predictive of toxicity than change in charge itself. For double mutant variants, we summed the feature values of the constituent singles for both AA property and aggregation/structure algorithm features.

Regression models were built using either (i) all variants, restricted to those occurring either (ii) inside or (iii) outside the toxicity hotspot (for double mutants both mutations have to occur either inside or outside the hotspot region), or (iv) including a binary location variable (0: one/all outside, 1: one inside, one outside, 2: one/all inside toxicity hotspot) and a third term indicating the interaction between location and the AA property or aggregation/structure algorithm feature.

**Predicting secondary structure from epistasis.** Epistasis is the non-independence of mutation effects, i.e., the toxicity of double mutants is different from that expected given the toxicity of their constituent single mutant variants. We have previously shown that epistasis between double mutants can result from structural interactions within proteins and therefore can be used to infer secondary and tertiary structural features.\(^{31,40}\)

In brief, double mutants were classified as epistatic if they had more extreme toxicity values (below 5th percentile or above 95th percentile) than other double mutants with similar single mutant toxicities, which was estimated from non-parametric surface fits of double mutant toxicity as a function of a two-dimensional set of toxicity space (Fig. 6a).

Double mutants close to the lower or upper measurement range limits (where the power to detect significant epistasis is reduced) were excluded from epistasis quantification. We calculated position-pair enrichments for epistatic double mutants resulting in a pair-wise enrichment matrix. 3D-geometrical features from this matrix were used as column-wise enrichments. An epistasis correlation score matrix was then derived from this enrichment matrix by calculating the partial correlation of epistasis interaction profiles (columns of the enrichment matrix) between all pairs of positions. The rationale for the correlation score is that structurally close positions within a protein should have similar epistatic interactions with all other positions in the protein. Calculating partial correlations additionally removes transitive effects and was found to be superior over epistasis enrichments in estimating secondary structures.\(^{41}\)

Secondary structure propensities were calculated by testing for agreement of epistasis correlation score patterns with the stereotypical periodicities of an α-helix and β-strand, using two-dimensional kernels at each position along the diagonal of the epistasis correlation score matrix.\(^{51}\)

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

The data that support the findings of this study are available from the corresponding author upon request. Raw sequencing data and the processed data table (Supplementary Data 3) have been deposited in NCBI’s Gene Expression Omnibus (GEO) and are accessible through the GEO Series accession number GSE128165. The source data underlying Fig. 3 and Supplementary Figs. 5 and 6 are provided as a Source Data file.

**Code availability**

All software code and custom scripts are available on GitHub: https://github.com/lehner-lab/DeDiSum for raw read processing, https://github.com/lehner-lab/tardipms for all downstream analyses and to produce all figures, and https://github.com/lehner-lab/tardipms-cellprofiler-scripts for CellProfiler pipelines.

Received: 20 May 2019 Accepted: 15 August 2019 Published online: 13 September 2019
References

1. Buratti, E. Functional Significance of TDP-43 Mutations in Disease. Adv. Neurol. 91, 1–53 (2015).

2. Ling, S.-C., Polymenidou, M. & Cleveland, D. W. Converging mechanisms in ALS and FTD: disrupted RNA and protein homeostasis. Neuron 79, 416–438 (2013).

3. Chiti, F., Stefani, M., Taddei, N., Ramponi, G. & Dobson, C. M. Rationalisation of why so many mutations of proteins on peptide and protein aggregation rates. Nature 424, 805–808 (2003).

4. Eisenberg, D. & Jucker, M. The amyloid state of proteins in human diseases. Cell 148, 1188–1203 (2012).

5. Chiti, F. & Dobson, C. M. Protein misfolding, functional amyloid, and human disease. Annu. Rev. Biochem. 75, 333–366 (2006).

6. Polymeropoulos, M. H. et al. Mutation in the alpha-synuclein gene identified in families with Parkinson’s disease. Science 276, 2045–2047 (1997).

7. Arnold, E. S. et al. ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43. Proc. Natl Acad. Sci. USA 110, E736–E745 (2013).

8. Taylor, J. P., Brown, R. H. & Cleveland, D. W. Decoding ALS: from genes to mechanism. Nature 539, 197–206 (2016).

9. Gordon, D. et al. Single-copy expression of an amyotrophic lateral sclerosis-linked TDP-43 mutation (M337V) in BAC transgenic mice leads to altered nuclear TDP-43. Proc. Natl Acad. Sci. USA 108, 16273–16278 (2011).

10. De Strooper, B. Lessons from a failed proteinopathy model: exploring the molecular determinants of TDP-43 toxicity and inclusion formation. J. Neuropathol. Exp. Neurol. 67, 1159–1165 (2008).

11. Amador-Ortiz, C. et al. TDP-43 immunoreactivity in hippocampal sclerosis and Alzheimer’s disease. Ann. Neurol. 61, 435–445 (2007).

12. Schwab, C., Arai, T., Hasegawa, M., Yu, S. & McGeer, P. L. Colocalization of transactivation-responsive DNA-binding protein 43 and Huntingtin in inclusions of Huntington disease. J. Neuropathol. Exp. Neurol. 67, 1159–1165 (2008).

13. Cremades, N. et al. Direct observation of the interconversion of normal and pathological TDP-43 conformations and implications for disease. Mol. Cell 38, 434–444 (2010).

14. McGurk, L. et al. Poly(ADP-Ribose) Prevents pathological phase separation of TDP-43. Nat. Neurosci. 21, 228–239 (2018).

15. Conicella, A. E., Zerze, G. H., Mittal, J. & Fawzi, N. L. ALS mutations disrupt phase separation mediated by a-helical structure in the TDP-43 low-complexity C-terminal domain. Struct. Fold. Des. 24, 1537–1549 (2016).

16. Halfmann, R. et al. Opposing effects of glutamine and asparagine govern protein aggregation rates. J. Biol. Chem. 288, 19614–19624 (2013).

17. Wegmann, S. et al. Tau protein liquid-liquid phase separation can initiate tau proteinopathy model: exploring the molecular determinants of TDP-43 aggregation and cellular toxicity. Proc. Natl Acad. Sci. USA 105, 6439–6444 (2008).

18. Bolognesi, B. et al. A concentration-dependent liquid phase separation can trigger alpha-synuclein amyloid formation in vitro. J. Biol. Chem. 288, 19614–19624 (2013).

19. Mateju, D. et al. An aberrant phase transition of stress granules triggered by a nuclear-import receptor cointeracting with TDP-43. J. Cell Biol. 218, 3661–3676 (2019).

20. Khan, T. et al. Quantifying nucleation in vivo reveals the physical basis of nucleic acid and protein aggregation rates. Cell 173, 677–692.e20 (2018).

21. Franzmann, T. M. et al. Phase separation of a yeast prion protein promotes cellular fitness. Science 359, 666543 (2018).

22. Wang, J. et al. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. Cell 174, 688–699.e16 (2018).

23. Patel, A. et al. A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease cell death. Cell 162, 1066–1077 (2015).

24. Moliex, A. et al. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. Cell 163, 123–133 (2015).

25. Wegrzyn, S. et al. Tau protein liquid-liquid phase separation can initiate tau aggregation. EMBO J. 37, e90049 (2018).

26. Vavouri, T., Semple, J. I., Garcia-Verdugo, R. & Lehner, B. Intrinsic protein disorder and interaction promiscuity are widely associated with disease sensitivity. Cell 158, 198–208 (2014).

27. Bolognesi, B. et al. A concentration-dependent liquid phase separation can trigger aggregation of protein expression. Cell Rep. 16, 222–231 (2016).

28. Halfmann, R. et al. Opposing effects of glutamine and asparagine govern protein formation by intrinsically disordered proteins. Mol. Cell 43, 72–84 (2011).

29. Halfmann, R. et al. Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. Science 314, 130–133 (2006).
73. Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively
Rapsomaniki, M. A. et al. easyFRAP: an interactive, easy-to-use tool for
the AXA Research Fund, the Bettencourt Schueller Foundation, and Agencia
grant (616434), the Spanish Ministry of Economy and Competitiveness (BFU2017-
89488-P), the EMBL Partnership, and the CERCA Program/Generalitat de Catalunya. We thank Pablo Baeza Centurión, Xavier Salvatella,
Alexandros Armaos and Benjamin Lang for discussion and assistance and the Eisenberg
lab for help with the ZipperDB analysis.

Author contributions
B.B. and B.L. conceived the project and designed the experiments; B.B. and M.S. per-
formed the experiments; A.J.F., B.B. and J.M.S. performed analyses of sequences; A.J.F.
and J.M.S. analysed the genetic interactions and structures; G.G.T. initiated, designed and
carried out the original computational analysis of physicochemical properties; B.B., A.J.F.
and B.L. wrote the manuscript with input from all authors.

Additional information
Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-
019-12101-z.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/
reprintsandpermissions/

Peer review information Nature Communications thanks the anonymous reviewers for
their contribution to the peer review of this work. Peer reviewer reports are available.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in
published maps and institutional affiliations.

Acknowledgements
Work in B.L.’s lab was supported by a European Research Council (ERC) Consolidator
grant (616434), the Spanish Ministry of Economy and Competitiveness (BFU2017-
89488-P), the AXA Research Fund, the Bettencourt Schueller Foundation, and Agencia
de Gestio d’Ajuts Universitari i de Recerca (AGAUR, SGR-831) G.G.T.’s lab was sup-
ported by the European Research Council (RIBOMYLOME_309545) and the Spanish
Ministry of Economy and Competitiveness (BFU2014-55054-P and BFU2017-86970-P).

© The Author(s) 2019