Missense mutations in intrinsically disordered protein regions link pathogenicity and phase separation

Short title: Missense mutations in phase-separating IDRs

Oliver L. Kipp,[1] Karen A. Lewis,[1] Loren E. Hough,[2,3] and Steven T. Whitten[1*]*

[1] Department of Chemistry and Biochemistry, Texas State University, San Marcos, TX, USA

[2] Department of Physics, University of Colorado Boulder, Boulder, CO, USA

[3] BioFrontiers Institute, University of Colorado Boulder, Boulder, CO, USA

* Co-corresponding authors

E-mail: loren.hough@colorado.edu

E-mail: steve.whitten@txstate.edu

Abstract

The impact of missense genetic variations on protein function is often enigmatic, especially for mutations that map to intrinsically disordered regions (IDRs). Given the functional importance of phase separation of IDRs, it has been proposed that mutations that modulate phase separation might preferentially lead to disease. To examine this idea, we used the robust predictability of phase-separating (PS) IDRs and annotation of disease-associated proteins and mutations to map the correlation between disease and phase separation. Consistent with previous work linking phase separation to cancer and autism spectrum disorder, we find a higher prevalence of predicted phase separation behavior in disease-associated proteins than typical for human proteins. We map the prevalence of phase separation across a wide range of diseases, finding that most, but not all, show an enrichment of phase separation in the proteins associated with them. Strikingly, the frequency of pathogenic mutations in predicted PS IDRs was elevated three-fold relative to IDRs not predicted to phase separate. Substitutions involving arginine and the aromatic types were among the most pathogenic for PS IDRs, while substitutions involving serine, threonine, and alanine the most benign. We applied these trends to mutations of uncertain clinical significance and predict that half found in PS IDRs are likely pathogenic. We find that phosphorylation sites were enriched in PS IDRs when compared to other protein regions, though mutations at such sites were mostly benign. Pathogenicity was highest for mutations in predicted PS IDRs when also found in a short linear motif, known mediators of protein-protein interactions.

Introduction

Large-scale sequencing efforts have resulted in catalogs of the human genome and its variations, providing insight on the genetic basis of traits and diseases (1–4). Genetic variations that cause single amino acid missense mutations are often found in the intrinsically disordered regions (IDRs) of proteins, as IDRs comprise approximately one-third of the human proteome (5, 6). IDR sequences are less conserved relative to the sequences found in protein regions that fold (7, 8). IDRs have been associated with key biological processes, including signal transduction (9, 10), chromatin remodeling (11), cell cycle control (12), and directing the curvature of membrane surfaces (13). IDRs also are essential mediators of liquid-liquid phase separation (14, 15), a phenomenon in which condensed liquid droplets separate from and subsequently coexist within bulk cytosol (16). Biological condensates such as nucleoli, Cajal bodies and stress granules are critical for intracellular compartmentalization and functional specificity (16–18). Human genetic variations that mutate IDR sequences thus potentially alter the function and properties of biological condensates, and possibly contribute to disease phenotypes (19).

Several studies provide evidence that dysregulation of protein phase separation has a role in human disease (20–22). Mittag and colleagues demonstrated that cancer-associated mutations in the tumor suppressor SPOP disrupt phase separation in vitro and interfere with the co-localization of SPOP with its substrates in nuclear speckles (23). Fusion oncoproteins, e.g., EWS-FLI1 (24, 25), spontaneously form condensates inside cells that promote cancer development (26, 27). Many proteins known to phase separate also have been found in neurodegenerative inclusions (22), for example α-synuclein in Parkinson's disease (28), tau in Alzheimer's disease (29), and TDP-43 (30) or FUS (31) in amyotrophic lateral sclerosis. Additionally, several predictors of phase separation have been used to identify correlations between phase separation and disease (32, 33), including cancer and autism spectrum disorder (19).

Previously, we developed an algorithm, ParSe, that accurately identifies from the protein primary sequence those IDRs likely to exhibit physiological phase separation behavior (34, 35). ParSe uses an optimal set of amino acid property scales for fast predictions of domain-level structure and provides a simple, quantitative metric for the sequence-calculated phase separation potential that obtains reasonable predictive power for existing mutant data (35). ParSe shows similar accuracy to other methods that have been developed to predict which protein sequences drive phase separation (35–41). One advantage of the ParSe algorithm is its speed, which enables it to identify drivers of phase separation at proteome scale (42).

Here, we investigate if ParSe can be used to identify relationships linking protein-mediated phase separation, human disease, and genetic missense variations. We found that human proteins associated with disease were enriched in predicted phase separation behavior relative to human proteins in general. Moreover, the incident rate of confirmed pathogenic missense mutations was three times higher in predicted phase-separating (PS) IDRs compared to IDRs not predicted to phase separate. This increase was not observed for confirmed benign missense mutations, suggesting the pathogenicity of mutations found in predicted PS IDRs was linked to the predicted phase separation behavior rather than some other IDR-driven function. Missense mutations at sites found in a short linear motif (SLiM) were especially pathogenic when also in a predicted PS IDR. In contrast, missense mutations at phosphorylation sites in a predicted PS IDR were primarily

benign. Comparing pathogenic and benign mutations in predicted PS IDRs revealed that single amino acid substitutions involving arginine and tyrosine were among the most pathogenic, which agrees with multiple reports showing the importance of arginine and tyrosine content in driving protein-mediated phase separation (38, 43–46). Collectively, our findings yield insight into the impact of genetic missense variations that mutate IDRs, revealing that pathogenicity associates with sequence features predicting IDR-mediated phase separation behavior.

Results

Construction of datasets

We first identified disease and comparison datasets in order to investigate the sequence-based differences between proteins in general and the subset that have been associated with disease, and between pathogenic missense mutations and benign missense mutations. The reference proteome for *Homo sapiens* (20,435 sequences) from UniProt (47) was used for the comparison set. Disease-associated proteins were obtained by searching UniProt for proteins annotated as "Human", "Reviewed (Swiss-Prot)", and the keyword "Disease". Disease-associated proteins (4,639 sequences) were sub-grouped according to condition, e.g., age-related macular degeneration, amyloidosis, asthma.

Confirmed pathogenic and benign missense mutations were obtained from two sources. First, we used the humsavar index of human variants curated from literature reports, which is part of the UniProt Knowledgebase (47). This index contained 32,665, 39,656, and 10,178 single amino acid substitutions that have been classified as likely pathogenic or pathogenic, likely benign or benign, or of uncertain clinical significance, respectively. We constructed a second missense mutation set from a natural variants index from UniProt. Mutations in this index were imported from the Ensembl Variation (48) and ClinVar databases (4) and include human variations from the 1,000 Genomes Project (3, 49). A small percentage (3%) of the single amino acid missense mutations in this index included predicted clinical significance; 59,435, 42,617, and 919,968 were predicted as likely pathogenic or pathogenic, likely benign or benign, or of uncertain clinical significance, respectively, and 35,019,669 single amino acid substitutions had no predicted clinical significance.

Confirmed SLiMs, which are short stretches of protein sequence that mediate protein-protein interactions, were obtained from the Eukaryotic Linear Motif (ELM) resource (50). Search of the ELM database for all "instances" filtered for "true positive" and "Homo sapiens" found 2,247 SLiMs from 1,371 human proteins. A curated index of 531,023 experimentally confirmed protein phosphorylation sites from 30,637 human proteins was obtained from the Eukaryotic Phosphorylation Site Database 2.0 (51).

Sequence-based calculation of phase separation potential

Dysregulation of protein phase separation has been associated with several human diseases (19–22). To explore this relationship broadly, we used the ParSe algorithm (35) to analyze the predicted prevalence of phase separation of disease-associated proteins as compared to the reference human proteome. ParSe predicts the domain-level organization of a protein from the primary sequence;

for folded, ID, and PS ID. The prediction is based on our finding of robust property differences between the three classes of protein regions (35). Briefly, we use a sliding-window approach to find where local properties in a sequence match the folded, ID, and PS ID classes. Figure 1 shows the ParSe prediction for the human protein FUS, confirmed to phase separate *in vitro* and in cells (52), that has a central, folded RNA-binding domain (53).

ParSe can also be used to calculate a predicted phase separation potential for individual proteins, taken as the sum of the classifier distance of each window matching the PS ID class. The classifier distance was developed to assess confidence in the window assignments (35). A higher "potential" can arise from longer predicted PS IDRs, a greater number of predicted PS IDRs, and higher confidence in these predictions. To enable a direct comparison between protein sets, we considered a cutoff in PS potential of 100. This value was chosen because every protein in the set of 43 confirmed to exhibit homotypic phase separation behavior, curated by Vernon et al (38), had a computed PS potential of at least 100 (Figure S1A). We use that minimum potential as a cut-off throughout this paper, identifying PS predicted proteins as those with a PS potential above 100.

Disease-associated protein sets are enriched in phase separation potential

ParSe predicts that human proteins annotated as disease-associated in UniProt have a higher prevalence of physiological phase separation when compared to the human proteome. Using a cutoff of PS potential of 100, ~16% of the human proteome and ~23% of the disease-associated set (a 44% increase) are likely to exhibit homotypic phase separation behavior (Figure S1A). Tsang et al, using the PScore algorithm (38), similarly found that disease-associated human proteins are enriched in phase separation potential (19).

As an alternative approach to quantify the differences in PS potential between protein sets, the percent of set data were used to create recall plots with the reference human proteome as the comparison set and from which the area under the curve (AUC) is calculated (Figure S1B). AUC values >0.5 indicate a sequence set enriched in computed phase separation potential relative to the reference human proteome. The disease-associated protein set reported an AUC of 0.563 (Table S1). Moreover, when sub-grouped by disease type, 79 of 100 disease types produced AUC >0.5 (Figure S2). In particular, this result recapitulates the previous finding that proteins associated with cancer and autism spectrum disorder have an over-representation of predicted phase-separating behavior (19).

We sought to determine the statistical significance of the calculated AUC values, complicated by the variability of the number of proteins annotated with each disease. We calculated the mean and standard deviation of the AUC determined from one hundred sets of proteins randomly sampled from the human proteome. Figure 2A shows the average AUC from random sets of human proteins, found to be tightly centered on 0.5, with the standard deviation depending on the number of proteins in the set. For example, for sets with fewer proteins (<10), one hundred random samples of the human proteome produced standard deviations of ~0.1 or more.

We found that the disease-associated protein set AUC was clearly a statistically significant difference from the reference human proteome. The AUC from the full set of disease-associated proteins, 0.563, was ~17 standard deviations higher than the average AUC from similarly sized

random protein sets (Table S1). Moreover, 51 of the disease sub-groups, out of 100, had AUC values exceeding the average plus the standard deviation from identically sized random sets (Figure 2A). Only 8 disease sub-groups had AUC values less than the average minus the standard deviation from the random protein sets.

We used the standard deviation to identify unusual and extreme AUC values, as three standard deviations from the mean cover 99.7% of data points in a normal distribution. The following disease sub-groups had computed AUC that were three standard deviations (or more) greater than the mean AUC of similarly sized random protein sets, in order of statistical significance: intellectual disability, proto-oncogene, deafness, epilepsy, autism spectrum disorder, tumor suppressor, neurodegeneration, epidermolysis bullosa, dwarfism, primary microcephaly, holoprosencephaly, craniosynostosis, palmoplantar keratoderma, Stickler syndrome, ectodermal dysplasia, Ehlers-Danlos syndrome, atrial septal defect, congenital hypothyroidism, and hypotrichosis (Figure 2B and Table S1). In comparison, disease sub-groups with computed AUC three standard deviations (or more) less than the mean AUC of similarly sized random protein sets were, in order of statistical significance: primary mitochondrial disease, Diamond-Blackfan anemia, and congenital disorder of glycosylation.

Distribution of missense variations by protein region class

Next, we considered the correlation between mutations (pathogenic, benign, or of unknown clinical significance) and phase separation. Proteins with confirmed pathogenic single amino acid missense mutations were enriched in predicted phase separation behavior (Figure 3A). We found that as compared to benign mutations, pathogenic mutations were more likely to be found in predicted PS proteins. 21.1% of human proteins listed in the humsavar index as harboring pathogenic missense mutations, as compared to 19.7% and 26.5% of human proteins with benign and uncertain clinical significance missense mutations, respectively, were predicted by ParSe to exhibit phase separation behavior, each an enrichment relative to the reference human proteome. The computed AUC values, with the reference human proteome as the comparison set, were 0.556, 0.540, and 0.604 for the protein sets with pathogenic, benign, and uncertain clinical significance missense mutations, respectively. Qualitatively similar results were obtained with the natural variants index, whereby proteins harboring single amino acid missense mutations were enriched in sequence-predicted phase separation behavior (Figure S3A).

Interestingly, both pathogenic and benign mutations were more likely to be found in PS predicted proteins than nonPS proteins (Figure 3B and Tables S2-S3). The incident rate for pathogenic missense mutations was calculated for two protein types: “predicted PS” (PS potential above 100, “PS”) and “predicted as not PS” (predicted PS potential below 100, “nonPS”)) and the three classes of protein region (predicted PS ID, nonPS ID, and folded). Incident rate by protein type was calculated as the number of proteins, PS or nonPS, with humsavar single amino acid pathogenic missense mutations divided by the total number of proteins with humsavar single amino acid pathogenic missense mutations, and then that ratio was normalized to the ratio in the human proteome for the protein type, PS or nonPS. The normalized incident rate for finding pathogenic missense mutations in predicted PS proteins was >1, and likewise for benign and uncertain clinical significance mutations, meaning that single amino acid missense mutations, regardless of clinical significance, were found at elevated rates in proteins with sequence features predicting phase

separation behavior, compared to human proteins in general. To investigate that finding further, we determined the percent composition of predicted ID residues in PS proteins as compared to nonPS proteins, establishing it was higher, 24% versus 10%. Thus, we hypothesize that the higher overall missense mutation rates can be explained by the higher ID content, consistent with previous observations (7, 8).

Strikingly, we found significant enrichment of pathogenic mutations within PS IDRs as compared to nonPS IDRs (Figure 3C). Protein region calculations of incident rates used the number of humsavar single amino acid pathogenic missense mutations found in a predicted region class divided by the total number of humsavar single amino acid pathogenic missense mutations, and then that ratio was normalized to the ratio of residues in the human proteome in a predicted region class divided by the total number of residues. Figure 3C shows that the normalized incident rate for pathogenic missense mutations in predicted PS IDRs (0.66) was approximately half the rate found in predicted folded regions (1.22). The PS IDR rate was almost 3-fold higher than the rate calculated in predicted nonPS IDRs (0.24). Thus, though the normalized incident rate for pathogenic missense mutations was highest in folded protein regions, pathogenic missense mutations in IDRs were primarily found in IDRs with sequence features predicting phase separation behavior. The odds ratio, which simplifies to the ratio of pathogenic to benign mutations in a region class, found that folded regions had a 2-fold higher pathogenicity than PS IDRs, and PS IDRs had a 3-fold higher pathogenicity than nonPS IDRs (Table S4).

In contrast, there was no significant difference in the normalized incident rate by protein region class for benign single amino acid missense mutations, found to be ~1 in predicted PS ID, nonPS ID, and folded protein regions (Figure 3C). Qualitatively similar results were obtained with the natural variants index, whereby the normalized incident rate for pathogenic missense mutations was highest in protein regions predicted to be folded, followed by PS IDRs, and then substantially lower in nonPS IDRs (Figure S3C).

Sequence dependence of missense variations by protein region class

We next sought to determine if pathogenic and benign single amino acid missense mutations exhibit substitution-specific preferences, and whether those preferences varied between the region classes. To achieve this, we compared the frequencies of amino acid specific substitutions for pathogenic versus benign and found significant differences between the types of mutations that tended to be pathogenic in predicted PS IDRs as compared to the other regions.

We first considered the entire set of mutations in the humsavar index (32,665 pathogenic mutations from 3,345 proteins) and found Arg-to-Cys, -His, -Gln, or -Trp, as well as Pro-to-Leu, Leu-to-Pro, Gly-to-Arg, and Glu-to-Lys as the most frequent pathogenic substitution types (Figure S4A). Many of these same substitutions (e.g., Arg-to-His, Arg-to-Gln, Pro-to-Leu, Glu-to-Lys) also were found in mutations of benign clinical significance (39,656 benign mutations from 11,682 proteins), emphasizing the importance of sequence context (Figure S4B). Figure 4A shows the difference in the calculated substitution frequencies, pathogenic minus benign, for all combinations of substitution types. Leu-to-Pro, Gly-to-Arg, and Arg-to-Trp had the highest net positive values (i.e., primarily pathogenic), while Val to Ile, Ile to Val, Ala-to-Thr, and Thr-to-Ala had the lowest net negative values (i.e., primarily benign).

Single amino acid missense mutations in regions identified by ParSe as matching the PS ID class reported different substitution preferences (Figure S5). Here, pathogenic mutations were strongly biased for substitutions from Gly, representing 40% (882 of 2208 substitutions) compared to 11% (496 of 4582 substitutions) of benign mutations. These pathogenic Gly substitutions were primarily from 27 collagen proteins (73%; 642 of 882). Substitutions from Gly would be destabilizing to the collagen triple helix, which structurally requires a repeating Gly-X-Y motif (54, 55). We removed the 27 collagen proteins from our analysis to reduce this bias for substitutions from Gly (removed 4%; 27 of 647 proteins with pathogenic mutations located in predicted PS regions). The resulting difference heat map of pathogenic and benign mutation frequencies (Figure 4B) shows Gly-to-Arg, Arg-to-Cys, and Cys-to-Arg as the most net pathogenic, while Thr-to-Ala, Ser-to-Asn, and Pro-to-Ser were the most net benign.

Mutations from or to an amino acid type were generally both net pathogenic or both net benign when found in predicted PS IDRs (Figure 5). This was calculated by summing the pathogenic versus benign difference frequencies for each amino acid type, across a row in Figure 4B to obtain the net variation from an amino acid type, and down a column to obtain the net variation to an amino acid type. Notable exceptions to the overall trend include Gly, which was strongly pathogenic for variations from Gly but not to Gly (even with collagen proteins removed), and Pro, which was often benign for variations from Pro but not to Pro. Single amino acid missense mutations involving Arg and the aromatics, Tyr, Trp, and Phe, were among the net pathogenic group (grey shading in Figure 5). Mutagenesis studies of proteins exhibiting homotypic phase separation behavior have found that Arg and the aromatic types, especially Tyr and Phe, contribute to cohesive protein-protein interactions that drive droplet formation (44–46). These findings are consistent with the idea that mutation at sites implicated in driving phase separation could lead to dysregulation of cellular processes that depend on phase separation, affecting human health. Single amino acid missense mutations involving Ser, Thr, and Ala, in that rank order, were strongly net benign.

For mutations found in regions matching the nonPS ID class, the types that tended to be benign were similar overall to the types that were benign in predicted PS IDRs. For example, substitutions involving Ala, Ser, Thr, Val, Met, and Ile were the most benign for both PS (Figure 5) and nonPS IDRs (Figure S6A-B). For pathogenic substitutions, this was generally the case too, where the types that tended to be pathogenic in nonPS or PS IDRs often involved the same group of amino acids, e.g., Arg, Cys, Tyr, Trp, and Gly. However, the specific substitutions that were strongly pathogenic in PS IDRs, as compared to nonPS, were different. In nonPS IDRs, Arg-to-Trp, Arg-to-Gln, and Glu-to-Lys were the most pathogenic. In PS IDRs, instead it was Gly-to-Arg, Arg-to-Cys, Cys-to-Arg, and Tyr-to-Cys. This is shown in Figure S6C by plotting the difference (Δ) of the difference frequency maps (PS - nonPS), where the more positive versus negative values identify substitutions with higher pathogenicity in PS IDRs versus nonPS, respectively. Notably, the difference in the net variation from or to an amino acid type (PS - nonPS), given in Figure S6D, strikingly shows the importance of sequence context regarding substitutions involving Arg, whereby mutations from Arg were significantly more benign, and mutations to Arg significantly more pathogenic in predicted PS versus nonPS IDRs. In rank order, mutations to Arg, from Gly, and from Cys exhibited the largest increases in pathogenicity when found in PS IDRs, as compared to nonPS, while mutations from Arg, to Trp, and from Thr showed the largest decreases.

For mutations found in protein regions matching the folded class, the types that tended to be benign or pathogenic again were similar to what was found in PS IDRs; e.g. mutations involving Arg, Cys, Tyr, Trp, and Gly tended to be pathogenic, while mutations involving Ala, Ser, Thr, Met, Val, and Ile tended to be benign (Figure S7A-B). We found the largest increases in pathogenicity for mutations in PS IDRs, relative to those found in predicted folded regions, for mutations involving Val and Ile, whereas the largest decreases in pathogenicity occurred for mutations involving Pro, Ser, and Leu (Figure S7C-D).

Predicting pathogenicity of missense variations by protein region class

The normalized incident rate for pathogenic missense mutations was found to depend strongly on protein region class (Figure 3C). Moreover, single amino acid substitution trends differed for pathogenic and benign missense mutations (Figures 4-5 and S5-S7). From these findings, we hypothesize that the pathogenic minus benign difference frequencies can be used to predict clinical significance from the substitution. While more sophisticated predictors of pathogenicity have been developed, typically focused on folded domains (56) or evolutionary models (57, 58), our goal was to leverage our observations of mutations in predicted PS IDRs to extrapolate the significance of the many missense mutations of unknown clinical significance in these regions, in particular.

This idea was tested using the difference frequency data in Figure 4A, finding that 65% of confirmed pathogenic single amino acid missense mutations were from a substitution type with a net positive difference frequency (i.e., net pathogenic), whereas 63% of confirmed benign single amino acid missense mutations were from a substitution type with a net negative difference frequency (Figure 6A). The combined true positive (correctly predicting pathogenic) and true negative (correctly predicting benign) rate was 64%. Thus, single amino acid missense substitutions were marginally predictive from difference frequency data when applied to the training set, with a 36% error rate. When these substitution trends from the humsavar index were used to predict clinical significance in the natural variance index, the combined true positive and true negative rate was similar, 62% (Figure S8).

For missense substitutions found in regions matching the PS ID class, the combined true positive and true negative rate was higher, 72% (Figure 6A), owing to a true negative rate that was markedly greater than the true positive rate (76% versus 64%). In protein regions matching the ID and folded classes, the combined true positive and true negative rates were 65% and 64%, respectively. In these calculations, difference frequencies were determined separately for each region class and the substitution-specific difference frequencies for a region class were used to predict clinical significance in that region class. Of note, when the contribution of collagen proteins was omitted from the calculated frequencies and excluded from the prediction tests, the observed predictability in PS IDRs for clinical significance was reduced to the same level found in the other protein region classes (Figure 6B). Thus, though pathogenicity was highest in folded protein regions for single amino acid missense mutations (Figure 3C), and ~3-fold higher in PS IDRs relative to nonPS IDRs, the overall predictability of clinical significance from region-specific substitution frequencies was mostly the same across the three region classes.

These substitution frequency trends predict that 51% and 41% of single amino acid missense mutations with uncertain clinical significance in the humsavar and natural variants indices, respectively, and 52% and 39% found within predicted PS IDRs, are pathogenic when excluding collagen proteins. 40% of the 35,019,669 single amino acid substitutions in the natural variants index with no predicted clinical significance, and 39% within predicted PS IDRs, also were predicted pathogenic. Despite the expected high error rate (~36%), the net of uncertain and unknown clinical significance missense mutations that were predicted to be pathogenic was substantial (~40-50%).

The high error rate in these predictions clearly suggests that clinical significance depends on additional input variables. Below, we explore two potential determinants of clinical significance that have been linked to IDR-mediated protein function.

SLiM and phosphorylation site distribution by region class and correlations to pathogenicity

SLiMs are conserved recognition motifs involved in signal transduction (59) and found primarily in the IDRs of the human proteome (60). Despite the functional importance of protein-protein interactions in phase separation, it is uncertain whether SLiMs play a role in phase separation, or if mutations in SLiMs identified in predicted PS proteins contribute to human diseases.

We first noticed that proteins containing SLiMs were enriched for phase separation. Using ParSe, we found that 38.1% of human proteins with confirmed SLiMs were predicted to exhibit phase separation behavior; an enrichment relative to the reference human proteome, where only 16.4% of human proteins are predicted to exhibit phase separation behavior (Figure 7A). The computed AUC of PS potential in proteins containing SLiMs was 0.688, significantly higher than 0.5 and by ~27 standard deviations as compared to similarly sized random protein sets (where $\sigma = 0.007$).

Consistent with a correlation between SLiMs and phase separation, the normalized incident rate for finding a SLiM in a predicted PS protein was substantially higher than the rate for nonPS proteins, 2.3 versus 0.74 (Figure 7B and Table S3). These rates were calculated as the number of proteins, PS or nonPS, found with a SLiM divided by the total number of SLiM containing proteins, and then that ratio was normalized to the reference proteome composition for PS and nonPS proteins. In contrast, among IDRs, confirmed SLiMs were enriched more so in nonPS than PS. Here, the normalized incident rates were calculated as the number of confirmed SLiMs found in a protein region class, divided by the total number of confirmed SLiMs and normalized to the ratio of the region class in the human proteome. The computed rates in PS ID, nonPS ID, and folded regions were 1.8, 2.5, and 0.42, respectively (Figure 7B and Table S3).

We next sought to quantify rates of missense mutations in SLiMs. Only 0.43% of confirmed pathogenic single amino acid missense mutations were found in a confirmed SLiM. However, because confirmed SLiMs were an even smaller fraction of the reference human proteome (0.14%), the computed normalized incident rate for pathogenic missense mutations in SLiMs was ~3. The normalized incident rate for confirmed benign and uncertain clinical significance missense mutations in a confirmed SLiM was 1.3 and 13, respectively (Table S3).

Pathogenic mutations were particularly enriched in SLiMs in PS IDRs, though the overall number of such mutations was small. The normalized incident rate of pathogenic SLiMs in PS IDRs was ~11, the highest of all pathogenic incident rates found in this work; 44 of 32,665 pathogenic missense mutations (0.13%) in the humsavar index were found in both a confirmed SLiM and a predicted PS IDR, whereas confirmed SLiMs in predicted PS IDRs represented 0.012% of the reference human proteome (Figure 7C). For comparison, the normalized incident rate was 1.6 for benign missense mutations found in both a confirmed SLiM and a predicted PS IDR. The pathogenic to benign odds ratio for missense mutations in a SLiM in a predicted PS IDR was 5.5 (Table S4), reflecting the significantly higher pathogenic normalized incident rate. The normalized incident rate for pathogenic mutation in both a SLiM and predicted nonPS IDR, and a SLiM and predicted folded region was 1.7 and 3.1, respectively (Figure 7C). Pathogenicity thus was highest, and by a significant margin, for missense mutations in a SLiM that also occurred in a predicted PS IDR.

Protein phosphorylation is used biologically as a mechanism to control many critical life processes (61, 62) and phosphorylation often occurs within protein regions that are ID (63). Based on our finding that mutations in PS IDRs involving Ser or Thr were statistically the least pathogenic (Figure 5), and considering that Ser and Thr were the most frequent sites of protein phosphorylation (Figure S9), we used ParSe to investigate the pathogenicity of mutations found at both a confirmed phosphorylation site and in a predicted PS IDR.

We found that sites of protein phosphorylation were preferentially found in predicted PS, as compared to nonPS, proteins (Figure 7A-B and Table S3). 18.8% of human proteins with confirmed phosphorylation sites were predicted by ParSe to exhibit phase separation behavior; an enrichment relative to the reference human proteome. The computed AUC of PS potential for proteins with a phosphorylation site was 0.525 (~90 standard deviations greater than random). The normalized incident rate for confirmed phosphorylation sites in predicted PS proteins and nonPS proteins was 1.2 and 0.97, respectively, while the normalized incident rate for predicted PS IDRs, nonPS IDRs, and folded regions was 2.7, 1.8, and 0.67, respectively. Thus, the normalized incident rate for protein phosphorylation was highest in predicted PS ID regions or domains.

However, though phosphorylation sites preferentially map to predicted PS IDRs, we found that missense mutations at such sites were overwhelmingly benign. 54 of 32,665 pathogenic missense mutations (0.17%) in the humsavar index were found both at a confirmed phosphorylation site and in a predicted PS IDR, yielding a normalized incident rate of 0.24 (Figure 7D and Table S3). The normalized incident rate for pathogenic missense mutations at a confirmed phosphorylation site and in a predicted nonPS IDR was 0.23, and in a predicted folded region was 0.99. The pathogenic to benign odds ratio was similarly low, 0.26 and 0.22, for phosphorylation sites in PS and nonPS IDRs, respectively, when compared to phosphorylation sites in folded regions (Table S4). Thus, pathogenicity was highest for missense mutations at a phosphorylation site that also occurred in a predicted folded region, and substantially more benign (~4-fold) for mutations at a phosphorylation site in predicted IDRs, both PS and nonPS. This is consistent with our finding that mutations involving Ser and Thr were statistically the most benign, especially when found in protein regions matching the PS ID class (Figure 5).

Discussion

It has been proposed that mutations that alter IDR-induced phase separation may have a higher likelihood of causing disease (19–22). Indeed, the results from several studies indicate a role for aberrant physiological protein phase separation in the progression of some human disorders (23–31). In this work, we identified disease and comparison datasets to map the prevalence of phase separation across a wide range of human diseases, finding that most show an enrichment for phase separation in the proteins associated with them. The diseases most strongly linked to phase separation included phenotypes associated with intellectual disability (e.g., autism, microcephaly, holoprosencephaly), cancer, neurodegeneration, and the skin (e.g. epidermolysis bullosa, palmoplantar keratoderma, Ehlers-Danlos syndrome). We also found that genetic missense variations, regardless of clinical significance, occur at higher rates in proteins predicted to exhibit phase separation behavior, which we attribute to higher ID content in predicted PS proteins.

Supporting a link between phase separation and disease, pathogenic missense mutations in predicted IDRs were found more frequently in PS as compared to nonPS regions, by a factor of approximately 3. Thus, although IDRs, and even PS IDRs, show low evolutionary conservation of sequence, mutations in these regions can be functionally significant. Using genetic variations from the humsavar index, we found 1186 pathogenic and 2191 benign missense mutations in predicted PS IDRs, compared to 727 pathogenic and 3828 benign missense mutations in predicted nonPS IDRs (Table S2). The pathogenic to benign odds ratio was highest for missense mutations in a predicted PS IDR when also in a confirmed SLiM. This combination gave an odds ratio of 5.5 that was significantly higher than the pathogenic to benign odds ratio for missense mutations in a SLiM and either of the other region classes (Table S4). This trend also was observed in the normalized incident rates for pathogenic missense mutations (Figure 7C). Based on these statistical correlations, we propose that SLiMs, in conjunction with PS IDRs, play a role in mediating the formation and composition of droplets in cells.

The clinical significance of missense mutations in predicted PS IDRs revealed a dependence on amino acid type. Single amino acid substitutions involving Cys, Arg, Gly, Asp, His, and the aromatics, Tyr, Phe, and Trp, were mostly pathogenic, whereas substitutions involving the other common amino acids were mostly benign (Figure 5). Cys residues are known to promote protein phase separation under oxidizing conditions (64). Numerous mutagenesis studies find Arg, Tyr, and Phe as mediators of droplet formation (38, 43–46). Mutating charge patterns through Asp substitution has been found to modulate phase separation potential (43). Protein-protein interactions involving His have been shown to contribute to phase separation behavior (65), especially toward pH responsive self-assembling processes (66). We conclude that this agreement between amino acid types known to facilitate protein phase separation and the pathogenicity of missense mutations found in predicted PS IDRs reflects the adverse effects of mutating residues that modulate the stability, function, and properties of biological condensates. This work also provides a foundation from which to optimize prediction of the effects of mutations on phase separation, which has historically been difficult.

The computed difference frequencies, pathogenic minus benign, by substitution type for missense mutations in predicted PS IDRs correctly assigned clinical significance to 72% of humsavar missense mutations residing in a predicted PS IDR, and 65% when excluding collagen proteins (Figure 6). From this, we conclude that the properties and functions of IDR-driven biological

condensates are tuned by the bulk composition of the corresponding PS IDRs. Our results highlight the efficacy of separately considering the role of mutation in PS IDRs as compared to other region types. Efforts to improve prediction of pathogenicity in these domains will likely require additional input variables that provide cellular context, e.g., localization, interacting partners, and gene ontology.

Methods

Protein Databases

The one-sequence-per-gene version of the reference human proteome, UP000005640, was downloaded from the Universal Protein Knowledgebase (47), UniProt, in November 2023. This proteome was used as the comparison protein set. Disease-associated human proteins were obtained by searching the UniProt database for proteins annotated as "Human", "Reviewed (Swiss-Prot)", and the keyword "Disease", also in November 2023. The proteins in this set were grouped by disease type by UniProt. Human proteins with confirmed SLiMs were found by search of the Eukaryotic Linear Motif database (50), in January 2024. Human proteins with confirmed phosphorylation sites were obtained from the Eukaryotic Phosphorylation Site Database 2.0 (51) in March 2024.

Genetic Variations Databases

The human variants index, file humsavar.txt release 2023_05 of 08-Nov-2023, and the natural variants index, file homo_sapiens_variation.txt release 2023_05 of 11-Oct-2023, both were downloaded from ftp.uniprot.org in November 2023.

Sequence-calculated Phase Separation Potential

Calculations of phase separation potential for individual protein sequences used the ParSe algorithm, version 2, described in full elsewhere (35) and available as a webtool (42) at stevewhitten.github.io/Parse_v2_web. Source code for both the algorithm and the webtool can be downloaded from github.com/stevewhitten. Calculations of phase separation potential for sequence sets in the FASTA format, yielding AUC and percent of set values reported in this study, used the webtool at stevewhitten.github.io/ParSe_v2_FASTA_reviewed_proteome with source code available at github.com/stevewhitten.

Calculation of Normalized Incident Rate

Normalized incident rates by protein type were calculated as the number of PS or non PS proteins found with single amino acid missense mutations in a variant index (humsavar or natural variants index) divided by the total number of proteins with single amino acid missense mutations in a variant index, and then dividing by the fraction of proteins in the reference human proteome of that type (PS or nonPS). Pathogenic rates were calculated using missense mutations labeled likely pathogenic or pathogenic; benign rates used missense mutations labeled likely benign or benign; uncertain clinical significance rates used missense mutations labeled uncertain clinical

significance. Protein type was determined by the ParSe algorithm, version 2, where PS proteins had PS potential ≥100.

Normalized incident rates by protein region class, for PS ID, nonPS ID, and folded, were calculated as the number of single amino acid missense mutations found in a region class divided by the total number of single amino acid missense mutations in a variant index, and then dividing by the fraction of residues in the reference human proteome for that region class. Pathogenic rates were calculated using missense mutations labeled likely pathogenic or pathogenic; benign rates used missense mutations labeled likely benign or benign; uncertain clinical significance rates used missense mutations labeled uncertain clinical significance. Protein region class by position in a primary sequence was determined using ParSe version 2.

Normalized incident rates for SLiMs and phosphorylation sites by protein type and region class were calculated identically as described above with the exceptions of using SLiM positions and phosphorylation sites, respectively, instead of single amino acid missense mutation sites.

Normalized incident rates for recursive conditions, e.g., pathogenic mutations in SLiMs in PS IDRs, were calculated as the fraction of mutations in a variant index satisfying the conditions divided by the fraction of residues the reference human proteome satisfying the non-mutation recursive conditions, e.g., SLiMs in PS IDRs.

Calculation of Odds Ratio

Odds ratios were used to compare the relative odds of an outcome. For pathogenic versus benign mutations, this was determined as the pathogenic odds divided by the benign odds. Pathogenic odds were calculated as the number of pathogenic mutations found in a protein type or region class divided by the number of residues in a protein type or region class. Benign odds were calculated as the number of benign mutations found in a protein type or region class divided by the number of residues in a protein type or region class. For a given protein type or region class, the odds ratio thus simplifies to the number of pathogenic mutations divided by the number of benign mutations found in a protein type or region class.

Calculation of Substitution Frequencies

Substitution frequencies were calculated as the fraction of single amino acid missense mutations of a given type, e.g., Gly-to-Ala, in a variant index. Substitution frequencies by region class, for PS ID, ID, and folded, were calculated as the fraction of single amino acid missense mutations of a given type, e.g., Gly-to-Ala, in a variant index for those found in a region class. Protein region class by position in a primary sequence was determined using ParSe version 2.

Additional Information

Supporting Information – This article contains supporting information (35, 47, 50, 51).

Author Contributions – L. E. H. and S. T. W. conceptualization; O. L. K., K. A. L., L. E. H., and S. T. W. formal analysis; O. L. K., K. A. L., L. E. H., and S. T. W. investigation; L. E. H. and S.

T. W. methodology; S. T. W. writing–original draft; O. L. K., K. A. L., L. E. H., and S. T. W. writing–review and editing.

Funding – This work was supported by the National Science Foundation under grant 1943488 (Loren E. Hough), the Sloan Foundation under grant G-2022-19553 (Karen A. Lewis), and Texas State University Office of Research and Sponsored Projects through the Research Enhancement Program (Steven T. Whitten and Karen A. Lewis). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF or Sloan Foundation.

Conflict of Interest – The authors declare that they have no conflicts of interest with the contents of this article.

Abbreviations – The abbreviations used are: AUC, area under the curve; ELM, Eukaryotic Linear Motif; ID, intrinsically disordered; IDR, intrinsically disordered region; nonPS, nonphase-separating; PS, phase-separating; SLiM, short linear motif; UniProt, Universal Protein Knowledgebase.

References

1. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F. A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A., Morgan, M. J., International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, C. for G. R., The Sanger Centre:, Washington University Genome Sequencing Center, US DOE Joint Genome Institute:, Baylor College of Medicine Human Genome Sequencing Center:, RIKEN Genomic Sciences Center:, Genoscope and CNRS UMR-8030:, Department of Genome Analysis, I. of M. B., GTC Sequencing Center:, Beijing Genomics Institute/Human Genome Center:, Multimegabase Sequencing Center, T. I. for S. B., Stanford Genome Technology Center:, University of Oklahoma's Advanced

Center for Genome Technology:, Max Planck Institute for Molecular Genetics:, Cold Spring Harbor Laboratory, L. A. H. G. C., GBF—German Research Centre for Biotechnology:, *Genome Analysis Group (listed in alphabetical order, also includes individuals listed under other headings):, Scientific management: National Human Genome Research Institute, U. N. I. of H., Stanford Human Genome Center:, University of Washington Genome Center:, Department of Molecular Biology, K. U. S. of M., University of Texas Southwestern Medical Center at Dallas:, Office of Science, U. D. of E., and The Wellcome Trust: (2001) Initial sequencing and analysis of the human genome. *Nature*. **409**, 860–921

2. 1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010) A map of human genome variation from population-scale sequencing. *Nature*. **467**, 1061–1073
3. Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., Marth, G. T., McVean, G. A., Nickerson, D. A., Schmidt, J. P., Sherry, S. T., Wang, J., Wilson, R. K., Gibbs, R. A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J. G., Zhu, Y., Wang, J., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Lander, E. S., Altshuler, D. M., Gabriel, S. B., Gupta, N., Gharani, N., Toji, L. H., Gerry, N. P., Resch, A. M., Flicek, P., Barker, J., Clarke, L., Gil, L., Hunt, S. E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R. E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Bentley, D. R., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Lehrach, H., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Borodina, T. A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.-L., Mardis, E. R., Wilson, R. K., Fulton, L., Fulton, R., Sherry, S. T., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., McVean, G. A., Durbin, R. M., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T. M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Schmidt, J. P., Davies, C. J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Auton, A., Campbell, C. L., Kong, Y., Marcketta, A., Gibbs, R. A., Yu, F., Antunes, L., Bainbridge, M., Muzny, D., Sabo, A., Huang, Z., Wang, J., Coin, L. J. M., Fang, L., Guo, X., Jin, X., Li, G., Li, Q., Li, Y., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Marth, G. T., Garrison, E. P., Kural, D., Lee, W.-P., Fung Leong, W., Stromberg, M., Ward, A. N., Wu, J., Zhang, M., Daly, M. J., DePristo, M. A., Handsaker, R. E., Altshuler, D. M., Banks, E., Bhatia, G., del Angel, G., Gabriel, S. B., Genovese, G., Gupta, N., Li, H., Kashin, S., Lander, E. S., McCarroll, S. A., Nemesh, J. C., Poplin, R. E., Yoon, S. C., Lihm, J., Makarov, V., Clark, A. G., Gottipati, S., Keinan, A., Rodriguez-Flores, J. L., Korbel, J. O., Rausch, T., Fritz, M. H., Stütz, A. M., Flicek, P., Beal, K., Clarke, L., Datta, A., Herrero, J., McLaren, W. M., Ritchie, G. R. S., Smith, R. E., Zerbino, D., Zheng-Bradley, X., Sabeti, P. C., Shlyakhter, I., Schaffner, S. F., Vitti, J., Cooper, D. N., Ball, E. V., Stenson, P. D.,

Bentley, D. R., Barnes, B., Bauer, M., Keira Cheetham, R., Cox, A., Eberle, M., Humphray, S., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E. E., Batzer, M. A., Konkel, M. K., Walker, J. A., MacArthur, D. G., Lek, M., Sudbrak, R., Amstislavskiy, V. S., Herwig, R., Mardis, E. R., Ding, L., Koboldt, D. C., Larson, D., Ye, K., Gravel, S., The 1000 Genomes Project Consortium, Corresponding authors, Steering committee, Production group, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, Coriell Institute for Medical Research, European Molecular Biology Laboratory, E. B. I., Illumina, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, US National Institutes of Health, University of Oxford, Wellcome Trust Sanger Institute, Analysis group, Affymetrix, Albert Einstein College of Medicine, Bilkent University, Boston College, Cold Spring Harbor Laboratory, Cornell University, European Molecular Biology Laboratory, Harvard University, Human Gene Mutation Database, Icahn School of Medicine at Mount Sinai, Louisiana State University, Massachusetts General Hospital, McGill University, and National Eye Institute, N. (2015) A global reference for human genetic variation. *Nature*. **526**, 68–74

4. Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W., and Maglott, D. R. (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*. **44**, D862-868
5. Oates, M. E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M. J., Xue, B., Dosztányi, Z., Uversky, V. N., Obradovic, Z., Kurgan, L., Dunker, A. K., and Gough, J. (2013) D2P2: database of disordered protein predictions. *Nucleic Acids Res*. **41**, D508–D516
6. van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D. T., Kim, P. M., Kriwacki, R. W., Oldfield, C. J., Pappu, R. V., Tompa, P., Uversky, V. N., Wright, P. E., and Babu, M. M. (2014) Classification of intrinsically disordered regions and proteins. *Chem Rev*. **114**, 6589–6631
7. Brown, C. J., Johnson, A. K., and Daughdrill, G. W. (2010) Comparing models of evolution for ordered and disordered proteins. *Mol Biol Evol*. **27**, 609–621
8. Light, S., Sagit, R., Sachenkova, O., Ekman, D., and Elofsson, A. (2013) Protein expansion is primarily due to indels in intrinsically disordered regions. *Mol Biol Evol*. **30**, 2645–2653
9. Xie, H., Vucetic, S., Iakoucheva, L. M., Oldfield, C. J., Dunker, A. K., Uversky, V. N., and Obradovic, Z. (2007) Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J Proteome Res*. **6**, 1882–1898
10. Liu, J., Perumal, N. B., Oldfield, C. J., Su, E. W., Uversky, V. N., and Dunker, A. K. (2006) Intrinsic disorder in transcription factors. *Biochemistry*. **45**, 6873–6888
11. El Hadidy, N., and Uversky, V. N. (2019) Intrinsic Disorder of the BAF Complex: Roles in Chromatin Remodeling and Disease Development. *Int J Mol Sci*. **20**, 5260
12. Cortese, M. S., Uversky, V. N., and Dunker, A. K. (2008) Intrinsic disorder in scaffold proteins: getting more from less. *Prog Biophys Mol Biol*. **98**, 85–106
13. Zeno, W. F., Thatte, A. S., Wang, L., Snead, W. T., Lafer, E. M., and Stachowiak, J. C. (2019) Molecular Mechanisms of Membrane Curvature Sensing by a Disordered Protein. *J. Am. Chem. Soc.* **141**, 10361–10371
14. Uversky, V. N., Kuznetsova, I. M., Turoverov, K. K., and Zaslavsky, B. (2015) Intrinsically disordered proteins as crucial constituents of cellular aqueous two phase systems and coacervates. *FEBS Letters*. **589**, 15–22

15. Gomes, E., and Shorter, J. (2019) The molecular language of membraneless organelles. *J. Biol. Chem.* **294**, 7115–7127
16. Brangwynne, C. P., Eckmann, C. R., Courson, D. S., Rybarska, A., Hoege, C., Gharakhani, J., Jülicher, F., and Hyman, A. A. (2009) Germline P Granules Are Liquid Droplets That Localize by Controlled Dissolution/Condensation. *Science*. **324**, 1729–1732
17. Brady, J. P., Farber, P. J., Sekhar, A., Lin, Y.-H., Huang, R., Bah, A., Nott, T. J., Chan, H. S., Baldwin, A. J., Forman-Kay, J. D., and Kay, L. E. (2017) Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E8194–E8203
18. Mitrea, D. M., and Kriwacki, R. W. (2016) Phase separation in biology; functional organization of a higher order. *Cell Communication and Signaling*. **14**, 1
19. Tsang, B., Pritišanac, I., Scherer, S. W., Moses, A. M., and Forman-Kay, J. D. (2020) Phase Separation as a Missing Mechanism for Interpretation of Disease Mutations. *Cell*. **183**, 1742–1756
20. Alberti, S., and Dormann, D. (2019) Liquid–Liquid Phase Separation in Disease. *Annual Review of Genetics*. **53**, 171–194
21. Wang, B., Zhang, L., Dai, T., Qin, Z., Lu, H., Zhang, L., and Zhou, F. (2021) Liquid–liquid phase separation in human health and diseases. *Sig Transduct Target Ther*. **6**, 1–16
22. Zbinden, A., Pérez-Berlanga, M., De Rossi, P., and Polymenidou, M. (2020) Phase Separation and Neurodegenerative Diseases: A Disturbance in the Force. *Dev Cell*. **55**, 45–68
23. Bouchard, J. J., Otero, J. H., Scott, D. C., Szulc, E., Martin, E. W., Sabri, N., Granata, D., Marzahn, M. R., Lindorff-Larsen, K., Salvatella, X., Schulman, B. A., and Mittag, T. (2018) Cancer Mutations of the Tumor Suppressor SPOP Disrupt the Formation of Active, Phase-Separated Compartments. *Mol Cell*. **72**, 19-36.e8
24. Ahmed, N. S., Harrell, L. M., Wieland, D. R., Lay, M. A., Thompson, V. F., and Schwartz, J. C. (2021) Fusion protein EWS-FLI1 is incorporated into a protein granule in cells. *RNA*. **27**, 920–932
25. Johnson, C. N., Sojitra, K. A., Sohn, E. J., Moreno-Romero, A. K., Baudin, A., Xu, X., Mittal, J., and Libich, D. S. (2024) Insights into Molecular Diversity within the FUS/EWS/TAF15 Protein Family: Unraveling Phase Separation of the N-Terminal Low-Complexity Domain from RNA-Binding Protein EWS. *J Am Chem Soc*. **146**, 8071–8085
26. Tripathi, S., Shirnekhi, H. K., Gorman, S. D., Chandra, B., Baggett, D. W., Park, C.-G., Somjee, R., Lang, B., Hosseini, S. M. H., Pioso, B. J., Li, Y., Iacobucci, I., Gao, Q., Edmonson, M. N., Rice, S. V., Zhou, X., Bollinger, J., Mitrea, D. M., White, M. R., McGrail, D. J., Jarosz, D. F., Yi, S. S., Babu, M. M., Mullighan, C. G., Zhang, J., Sahni, N., and Kriwacki, R. W. (2023) Defining the condensate landscape of fusion oncoproteins. *Nat Commun*. **14**, 6008
27. Shirnekhi, H. K., Chandra, B., and Kriwacki, R. W. (2023) The Role of Phase-Separated Condensates in Fusion Oncoprotein–Driven Cancers. *Annual Review of Cancer Biology*. **7**, 73–91
28. Ray, S., Singh, N., Kumar, R., Patel, K., Pandey, S., Datta, D., Mahato, J., Panigrahi, R., Navalkar, A., Mehra, S., Gadhe, L., Chatterjee, D., Sawner, A. S., Maiti, S., Bhatia, S., Gerez, J. A., Chowdhury, A., Kumar, A., Padinhateeri, R., Riek, R., Krishnamoorthy, G., and Maji, S. K. (2020) α-Synuclein aggregation nucleates through liquid-liquid phase separation. *Nat Chem*. **12**, 705–716

29. Ambadipudi, S., Biernat, J., Riedel, D., Mandelkow, E., and Zweckstetter, M. (2017) Liquid–liquid phase separation of the microtubule-binding repeats of the Alzheimer-related protein Tau. *Nat Commun*. 10.1038/s41467-017-00480-0
30. Prasad, A., Bharathi, V., Sivalingam, V., Girdhar, A., and Patel, B. K. (2019) Molecular Mechanisms of TDP-43 Misfolding and Pathology in Amyotrophic Lateral Sclerosis. *Front Mol Neurosci*. **12**, 25
31. Murthy, A. C., Dignon, G. L., Kan, Y., Zerze, G. H., Parekh, S. H., Mittal, J., and Fawzi, N. L. (2019) Molecular interactions underlying liquid-liquid phase separation of the FUS low-complexity domain. *Nat Struct Mol Biol*. **26**, 637–648
32. Sun, J., Qu, J., Zhao, C., Zhang, X., Liu, X., Wang, J., Wei, C., Liu, X., Wang, M., Zeng, P., Tang, X., Ling, X., Qing, L., Jiang, S., Chen, J., Chen, T. S. R., Kuang, Y., Gao, J., Zeng, X., Huang, D., Yuan, Y., Fan, L., Yu, H., and Ding, J. (2024) Precise prediction of phase-separation key residues by machine learning. *Nat Commun*. **15**, 2662
33. Feng, M., Wei, X., Zheng, X., Liu, L., Lin, L., Xia, M., He, G., Shi, Y., and Lu, Q. (2024) Decoding Missense Variants by Incorporating Phase Separation via Machine Learning. *Nat Commun*. **15**, 8279
34. Paiz, E. A., Allen, J. H., Correia, J. J., Fitzkee, N. C., Hough, L. E., and Whitten, S. T. (2021) Beta turn propensity and a model polymer scaling exponent identify intrinsically disordered phase-separating proteins. *Journal of Biological Chemistry*. **297**, 101343
35. Ibrahim, A. Y., Khaodeuanepheng, N. P., Amarasekara, D. L., Correia, J. J., Lewis, K. A., Fitzkee, N. C., Hough, L. E., and Whitten, S. T. (2023) Intrinsically disordered regions that drive phase separation form a robustly distinct protein class. *J Biol Chem*. **299**, 102801
36. Pancsa, R., Vranken, W., and Mészáros, B. (2021) Computational resources for identifying and describing proteins driving liquid-liquid phase separation. *Brief Bioinform*. 10.1093/bib/bbaa408
37. Vernon, R. M., and Forman-Kay, J. D. (2019) First-generation predictors of biological protein phase separation. *Current Opinion in Structural Biology*. **58**, 88–96
38. Vernon, R. M., Chong, P. A., Tsang, B., Kim, T. H., Bah, A., Farber, P., Lin, H., and Forman-Kay, J. D. (2018) Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *Elife*. 10.7554/eLife.31486
39. Cai, H., Vernon, R. M., and Forman-Kay, J. D. (2022) An Interpretable Machine-Learning Algorithm to Predict Disordered Protein Phase Separation Based on Biophysical Interactions. *Biomolecules*. **12**, 1131
40. Chu, X., Sun, T., Li, Q., Xu, Y., Zhang, Z., Lai, L., and Pei, J. (2022) Prediction of liquid-liquid phase separating proteins using machine learning. *BMC Bioinformatics*. **23**, 72
41. Orlando, G., Raimondi, D., Tabaro, F., Codicè, F., Moreau, Y., and Vranken, W. F. (2019) Computational identification of prion-like RNA-binding proteins that form liquid phase-separated condensates. *Bioinformatics*. **35**, 4617–4623
42. Wilson, C., Lewis, K. A., Fitzkee, N. C., Hough, L. E., and Whitten, S. T. (2023) ParSe 2.0: A web tool to identify drivers of protein phase separation at the proteome level. *Protein Sci*. **32**, e4756
43. Wang, J., Choi, J.-M., Holehouse, A. S., Lee, H. O., Zhang, X., Jahnel, M., Maharana, S., Lemaitre, R., Pozniakovsky, A., Drechsel, D., Poser, I., Pappu, R. V., Alberti, S., and Hyman, A. A. (2018) A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell*. **174**, 688-699.e16

44. Martin, E. W., Holehouse, A. S., Peran, I., Farag, M., Incicco, J. J., Bremer, A., Grace, C. R., Soranno, A., Pappu, R. V., and Mittag, T. (2020) Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science*. **367**, 694–699
45. Schuster, B. S., Dignon, G. L., Tang, W. S., Kelley, F. M., Ranganath, A. K., Jahnke, C. N., Simpkins, A. G., Regy, R. M., Hammer, D. A., Good, M. C., and Mittal, J. (2020) Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior. *Proc Natl Acad Sci U S A*. **117**, 11421–11431
46. Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., and Mittag, T. (2022) Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. *Nat. Chem.* **14**, 196–207
47. The UniProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*. **51**, D523–D531
48. Harrison, P. W., Amode, M. R., Austine-Orimoloye, O., Azov, A. G., Barba, M., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S. K., Boddu, S., Branco Lins, P. R., Brooks, L., Ramaraju, S. B., Campbell, L. I., Martinez, M. C., Charkhchi, M., Chougule, K., Cockburn, A., Davidson, C., De Silva, N. H., Dodiya, K., Donaldson, S., El Houdaigui, B., Naboulsi, T. E., Fatima, R., Giron, C. G., Genez, T., Grigoriadis, D., Ghattaoraya, G. S., Martinez, J. G., Gurbich, T. A., Hardy, M., Hollis, Z., Hourlier, T., Hunt, T., Kay, M., Kaykala, V., Le, T., Lemos, D., Lodha, D., Marques-Coelho, D., Maslen, G., Merino, G. A., Mirabueno, L. P., Mushtaq, A., Hossain, S. N., Ogeh, D. N., Sakthivel, M. P., Parker, A., Perry, M., Piližota, I., Poppleton, D., Prosovetskaia, I., Raj, S., Pérez-Silva, J. G., Salam, A. I. A., Saraf, S., Saraiva-Agostinho, N., Sheppard, D., Sinha, S., Sipos, B., Sitnik, V., Stark, W., Steed, E., Suner, M.-M., Surapaneni, L., Sutinen, K., Tricomi, F. F., Urbina-Gómez, D., Veidenberg, A., Walsh, T. A., Ware, D., Wass, E., Willhoft, N. L., Allen, J., Alvarez-Jarreta, J., Chakiachvili, M., Flint, B., Giorgetti, S., Haggerty, L., Ilsley, G. R., Keatley, J., Loveland, J. E., Moore, B., Mudge, J. M., Naamati, G., Tate, J., Trevanion, S. J., Winterbottom, A., Frankish, A., Hunt, S. E., Cunningham, F., Dyer, S., Finn, R. D., Martin, F. J., and Yates, A. D. (2024) Ensembl 2024. *Nucleic Acids Research*. **52**, D891–D899
49. Durbin, R. M., Altshuler, D., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S. B., Gibbs, R. A., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., McVean, G. A., Nickerson, D. A., Peltonen, L., Schafer, A. J., Sherry, S. T., Wang, J., Wilson, R. K., Gibbs, R. A., Deiros, D., Metzker, M., Muzny, D., Reid, J., Wheeler, D., Wang, J., Li, J., Jian, M., Li, G., Li, R., Liang, H., Tian, G., Wang, B., Wang, J., Wang, W., Yang, H., Zhang, X., Zheng, H., Lander, E. S., Altshuler, D., Ambrogio, L., Bloom, T., Cibulskis, K., Fennell, T. J., Gabriel, S. B., Jaffe, D. B., Shefler, E., Sougnez, C. L., Bentley, D. R., Gormley, N., Humphray, S., Kingsbury, Z., Kokko-Gonzales, P., Stone, J., McKernan, K. J., Costa, G. L., Ichikawa, J. K., Lee, C. C., Sudbrak, R., Lehrach, H., Borodina, T. A., Dahl, A., Davydov, A. N., Marquardt, P., Mertes, F., Nietfeld, W., Rosenstiel, P., Schreiber, S., Soldatov, A. V., Timmermann, B., Tolzmann, M., Egholm, M., Affourtit, J., Ashworth, D., Attiya, S., Bachorski, M., Buglione, E., Burke, A., Caprio, A., Celone, C., Clark, S., Conners, D., Desany, B., Gu, L., Guccione, L., Kao, K., Kebbel, A., Knowlton, J., Labrecque, M., McDade, L., Mealmaker, C., Minderman, M., Nawrocki, A., Niazi, F., Pareja, K., Ramenani, R., Riches, D., Song, W., Turcotte, C., Wang, S., Mardis, E. R., Wilson, R. K., Dooling, D., Fulton, L., Fulton, R., Weinstock, G., Durbin, R.

M., Burton, J., Carter, D. M., Churcher, C., Coffey, A., Cox, A., Palotie, A., Quail, M., Skelly, T., Stalker, J., Swerdlow, H. P., Turner, D., De Witte, A., Giles, S., Gibbs, R. A., Wheeler, D., Bainbridge, M., Challis, D., Sabo, A., Yu, F., Yu, J., Wang, J., Fang, X., Guo, X., Li, R., Li, Y., Luo, R., Tai, S., Wu, H., Zheng, H., Zheng, X., Zhou, Y., Li, G., Wang, J., Yang, H., Marth, G. T., Garrison, E. P., Huang, W., Indap, A., Kural, D., Lee, W.-P., Fung Leong, W., Quinlan, A. R., Stewart, C., Stromberg, M. P., Ward, A. N., Wu, J., Lee, C., Mills, R. E., Shi, X., Daly, M. J., DePristo, M. A., Altshuler, D., Ball, A. D., Banks, E., Bloom, T., Browning, B. L., Cibulskis, K., Fennell, T. J., Garimella, K. V., Grossman, S. R., Handsaker, R. E., Hanna, M., Hartl, C., Jaffe, D. B., Kernytsky, A. M., Korn, J. M., Li, H., Maguire, J. R., McCarroll, S. A., McKenna, A., Nemesh, J. C., Philippakis, A. A., Poplin, R. E., Price, A., Rivas, M. A., Sabeti, P. C., Schaffner, S. F., Shefler, E., Shlyakhter, I. A., Cooper, D. N., Ball, E. V., Mort, M., Phillips, A. D., Stenson, P. D., Sebat, J., Makarov, V., Ye, K., Yoon, S. C., Bustamante, C. D., Clarke, L., Flicek, P., Cunningham, F., Herrero, J., Keenen, S., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Smith, R. E., Zalunin, V., Zheng-Bradley, X., Korbel, J. O., Stütz, A. M., Humphray, S., Bauer, M., Keira Cheetham, R., Cox, T., Eberle, M., James, T., Kahn, S., Murray, L., Chakravarti, A., Ye, K., De La Vega, F. M., Fu, Y., Hyland, F. C. L., Manning, J. M., McLaughlin, S. F., Peckham, H. E., Sakarya, O., Sun, Y. A., Tsung, E. F., Batzer, M. A., Konkel, M. K., Walker, J. A., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Herwig, R., Parkhomchuk, D. V., Sherry, S. T., Agarwala, R., Khouri, H. M., Morgulis, A. O., Paschall, J. E., Phan, L. D., Rotmistrovsky, K. E., Sanders, R. D., Shumway, M. F., Xiao, C., McVean, G. A., Auton, A., Iqbal, Z., Lunter, G., Marchini, J. L., Moutsianas, L., Myers, S., Tumian, A., Desany, B., Knight, J., Winer, R., Craig, D. W., Beckstrom-Sternberg, S. M., Christoforides, A., The 1000 Genomes Project Consortium, Corresponding author, Steering committee, Production group: Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, Illumina, Life Technologies, Max Planck Institute for Molecular Genetics, Roche Applied Science, Washington University in St Louis, Wellcome Trust Sanger Institute, Analysis group: Agilent Technologies, Baylor College of Medicine, Boston College, Brigham and Women's Hospital, Cardiff University, T. H. G. M. D., Cold Spring Harbor Laboratory, Cornell and Stanford Universities, European Bioinformatics Institute, European Molecular Biology Laboratory, Johns Hopkins University, Leiden University Medical Center, Louisiana State University, US National Institutes of Health, Oxford University, and The Translational Genomics Research Institute (2010) A map of human genome variation from population-scale sequencing. *Nature*. **467**, 1061–1073

50. Kumar, M., Michael, S., Alvarado-Valverde, J., Zeke, A., Lazar, T., Glavina, J., Nagy-Kanta, E., Donagh, J. M., Kalman, Z. E., Pascarelli, S., Palopoli, N., Dobson, L., Suarez, C. F., Van Roey, K., Krystkowiak, I., Griffin, J. E., Nagpal, A., Bhardwaj, R., Diella, F., Mészáros, B., Dean, K., Davey, N. E., Pancsa, R., Chemes, L. B., and Gibson, T. J. (2024) ELM—the Eukaryotic Linear Motif resource—2024 update. *Nucleic Acids Research*. **52**, D442–D455
51. Lin, S., Wang, C., Zhou, J., Shi, Y., Ruan, C., Tu, Y., Yao, L., Peng, D., and Xue, Y. (2021) EPSD: a well-annotated data resource of protein phosphorylation sites in eukaryotes. *Brief Bioinform*. **22**, 298–307
52. Patel, A., Lee, H. O., Jawerth, L., Maharana, S., Jahnel, M., Hein, M. Y., Stoynov, S., Mahamid, J., Saha, S., Franzmann, T. M., Pozniakovski, A., Poser, I., Maghelli, N., Royer, L. A., Weigert, M., Myers, E. W., Grill, S., Drechsel, D., Hyman, A. A., and Alberti, S.

(2015) A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell*. **162**, 1066–1077
53. Liu, X., Niu, C., Ren, J., Zhang, J., Xie, X., Zhu, H., Feng, W., and Gong, W. (2013) The RRM domain of human fused in sarcoma protein reveals a non-canonical nucleic acid binding site. *Biochim Biophys Acta*. **1832**, 375–385
54. Bella, J., Eaton, M., Brodsky, B., and Berman, H. M. (1994) Crystal and Molecular Structure of a Collagen-Like Peptide at 1.9 Å Resolution. *Science*. **266**, 75–81
55. Ramshaw, J. A. M., Shah, N. K., and Brodsky, B. (1998) Gly-X-Y Tripeptide Frequencies in Collagen: A Context for Host–Guest Triple-Helical Peptides. *Journal of Structural Biology*. **122**, 86–91
56. Chasman, D., and Adams, R. M. (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J Mol Biol*. **307**, 683–706
57. Reva, B., Antipin, Y., and Sander, C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res*. **39**, e118
58. Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y., and Marks, D. S. (2021) Disease variant prediction with deep generative models of evolutionary data. *Nature*. **599**, 91–95
59. Diella, F., Haslam, N., Chica, C., Budd, A., Michael, S., Brown, N. P., Trave, G., and Gibson, T. J. (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front Biosci*. **13**, 6580–6603
60. Ren, S., Uversky, V. N., Chen, Z., Dunker, A. K., and Obradovic, Z. (2008) Short Linear Motifs recognized by SH2, SH3 and Ser/Thr Kinase domains are conserved in disordered protein regions. *BMC Genomics*. **9 Suppl 2**, S26
61. Cohen, P. (2001) The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture. *Eur J Biochem*. **268**, 5001–5010
62. Whitmarsh, A. J., and Davis, R. J. (2000) Regulation of transcription factor function by phosphorylation. *Cell Mol Life Sci*. **57**, 1172–1183
63. Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., and Dunker, A. K. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*. **32**, 1037–1049
64. Mondal, M., Jankoski, P. E., Lee, L. D., Dinakarapandian, D. M., Chiu, T.-Y., Swetman, W. S., Wu, H., Paravastu, A. K., Clemons, T. D., and Rangachari, V. (2024) Reversible Disulfide Bond Cross-Links as Tunable Levers of Phase Separation in Designer Biomolecular Condensates. *J Am Chem Soc*. **146**, 25299–25311
65. Gabryelczyk, B., Cai, H., Shi, X., Sun, Y., Swinkels, P. J. M., Salentinig, S., Pervushin, K., and Miserez, A. (2019) Hydrogen bond guidance and aromatic stacking drive liquid-liquid phase separation of intrinsically disordered histidine-rich peptides. *Nat Commun*. **10**, 5465
66. Kathe, N. C., Novakovic, M., and Allain, F. H.-T. (2024) Buffer choice and pH strongly influence phase separation of SARS-CoV-2 nucleocapsid with RNA. *Mol Biol Cell*. **35**, ar73

Figures

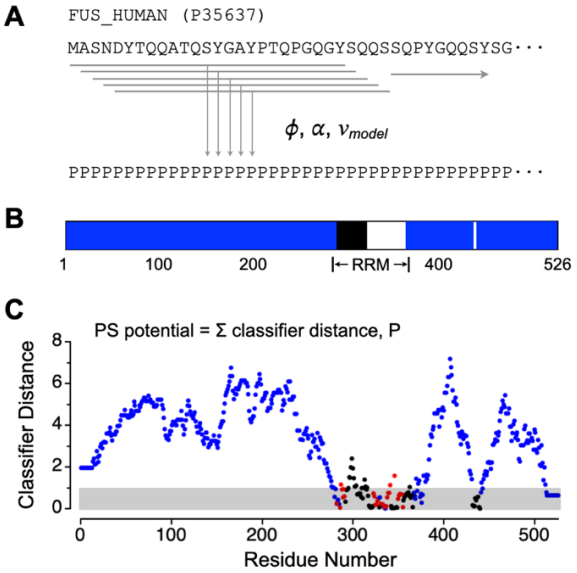

Figure 1. The ParSe algorithm (version 2). (**A**) A sliding window approach is used to identify regions within a protein that match the folded, ID, and PS ID classes. Hydrophobicity (ϕ), α-helix propensity (α), and v_{model} are calculated for each contiguous stretch of 25-residues, or "window", in the primary sequence. Each window is assigned a label, F, D, or P, depending on the values of ϕ, α, and v_{model}. The sequence properties of F-labeled windows match the folded class, D-labeled windows match the ID class, while P-labeled windows match the PS ID class. Window labels are assigned to the central residue of the window. Terminal residues are assigned labels according to the first and last windows. (**B**) Contiguous regions of at least 20 residues that are 90% of only one label P, D, or F are colored blue, red, or black, respectively, to represent predicted PS IDRs, IDRs, or folded regions. White corresponds to regions with a mixture of P, D, or F labels. (**C**) Classifier distance of each window, assigned to the central residue of the window and colored according to its label P (blue), D (red), or F (black). Values less than 1 in the classifier distance (shaded grey) indicate uncertainty for an assigned label. Panels A-C show results from ParSe for the primary sequence of the FUS RNA-binding protein that is known to phase separate and has a folded RNA Recognition Motif (RRM) at residues 288-368.

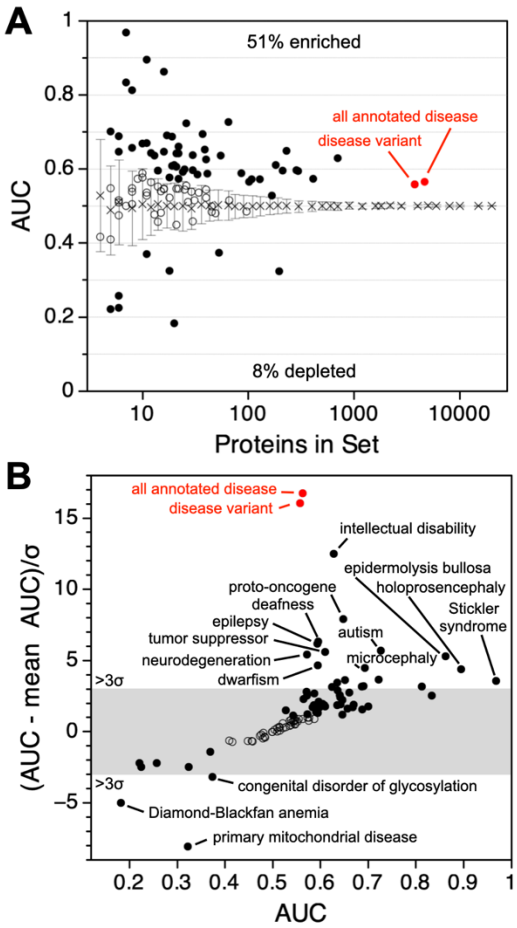

Figure 2. Disease-associated proteins are enriched in sequence predicted PS potential. (**A**) AUC was calculated for sets of proteins annotated as disease associated in UniProt and plotted according to the number of proteins in each set (open and filled circles). The reference human proteome was used as the comparison set to calculate AUC. The mean and standard deviation (x symbol with error bars) from random protein sets were determined by sampling the reference human proteome one hundred times at each indicated set size. Red circles show AUC for the full set of disease-associated proteins and the set containing all proteins for which at least one genetic variant involved in a disease has been reported. (**B**) Extreme AUC values were identified by subtracting from the AUC for a disease set (AUC) the mean AUC from random sets (mean AUC) and then dividing by the AUC standard deviation (σ) from the random sets. Panel B symbols and colors match panel A definitions.

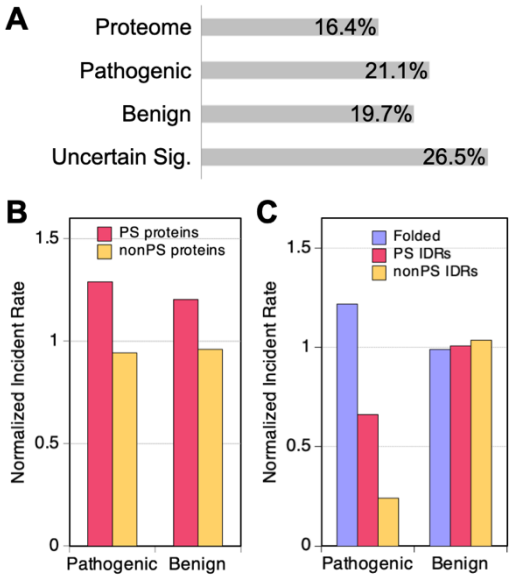

Figure 3. Distribution of missense mutations in the humsavar index. (**A**) Percent of proteins predicted as phase-separating by ParSe within the reference human proteome and within the subset of proteins containing pathogenic, benign, or uncertain clinical significance single amino acid missense mutations. Predicted phase-separating proteins had computed PS potentials ≥100. (**B**) Incident rate was calculated as the number of predicted PS (red) or nonPS (orange) proteins with missense mutations of a given type (pathogenic or benign) divided by the total number of proteins with mutations of that type, and then normalized to the reference human proteome composition, or (**C**) the number of missense mutations of a given type (pathogenic or benign) found in a region class (folded, PS ID, or nonPS ID), divided by the total number of mutations of a type and then normalized to the reference human proteome composition.

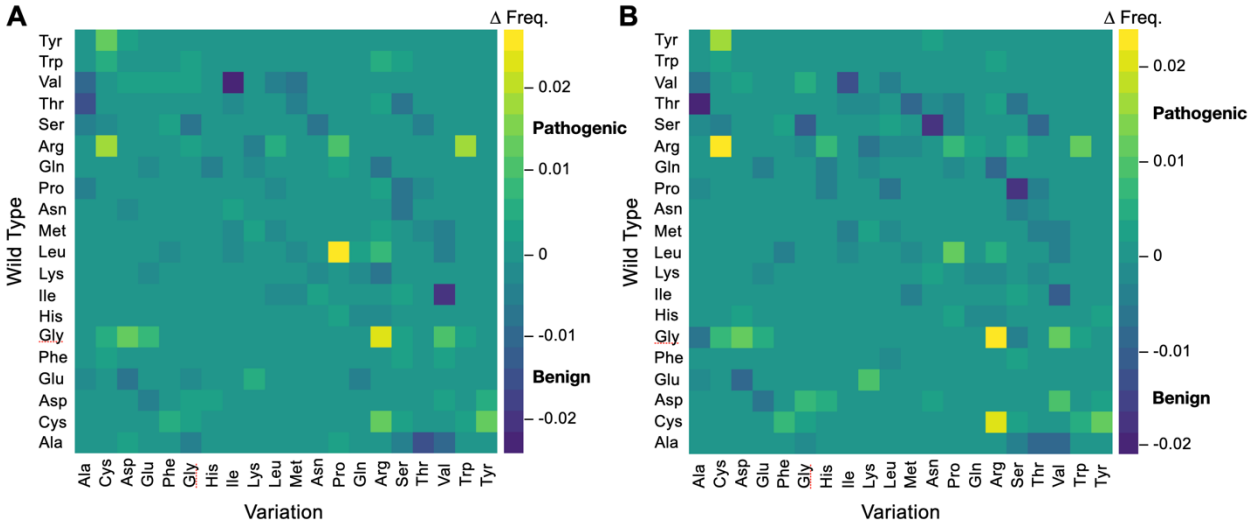

Figure 4. Difference frequencies of pathogenic minus benign single amino acid missense mutations. (**A**) All single amino acid missense mutations in the humsavar index, and (**B**) only those found in a protein region matching the PS ID class and excluding collagen proteins. Difference frequencies (Δ Freq., pathogenic rate minus benign rate) were colored according to the scale on the right of each plot.

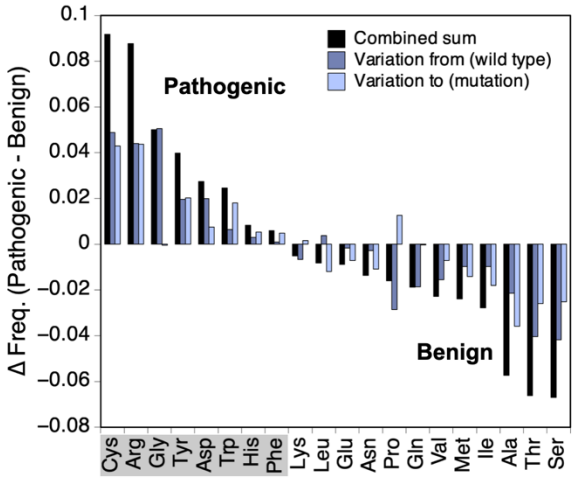

Figure 5. Pathogenicity by amino acid substitution in PS IDRs. Pathogenic minus benign difference frequencies (Δ Freq.) for positions in predicted PS IDRs were summed by wild type amino acid (blue) and mutation amino acid (light blue). The combined sum (black) is the simple addition of the wild type and mutation values. Contributions from collagen proteins were omitted.

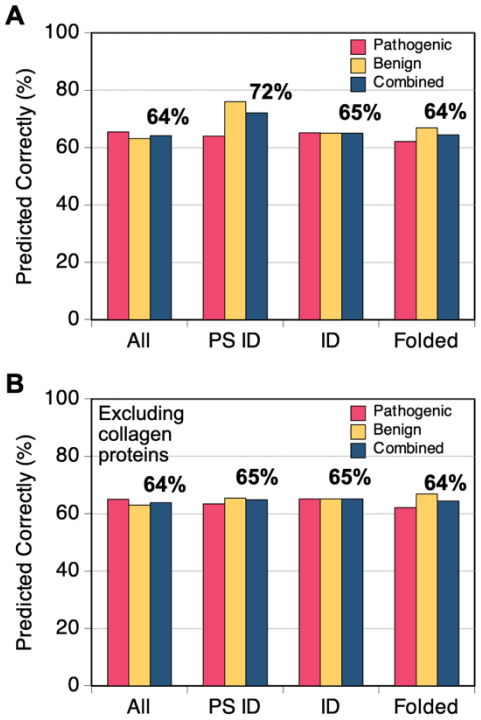

Figure 6. Predicting clinical significance of single amino acid missense mutations. (**A**) Pathogenic minus benign difference frequencies were calculated for all single amino acid missense mutations in the humsavar index and for the subsets residing in protein regions matching the PS ID, ID (i.e., nonPS ID), or folded classes. Clinical significance was predicted for missense mutations with known clinical significance where a positive difference frequency predicts pathogenic, and a negative difference frequency predicts benign. Shown is the percent predicted correctly. (**B**) The effect from excluding collagen proteins.

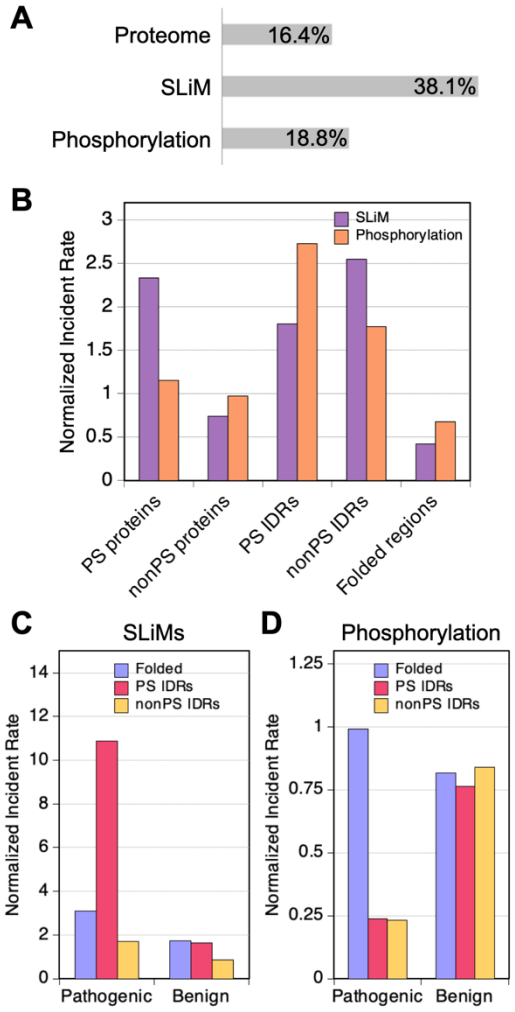

Figure 7. Distribution of SLiMs and phosphorylation sites by protein type, region class, and correlations to pathogenic and benign missense mutations. (**A**) Percent of proteins predicted as phase-separating by ParSe within the reference human proteome and within the subset of proteins containing a SLiM or phosphorylation site. Predicted phase-separating proteins had computed PS potentials ≥100. (**B**) Incident rates for finding a SLiM (purple) or phosphorylation site (salmon) in a protein type were calculated as the number of PS or nonPS proteins containing a SLiM or phosphorylation site (separately), divided by the total number of proteins with a SLiM or phosphorylation site, and then normalized to the reference human proteome composition for PS or nonPS proteins. Rates by region class were calculated as the number of SLiMs or phosphorylation sites found in a region class, divided by the total number of SLiMs or phosphorylation sites, and then normalized to the reference human proteome composition for a region class. Pathogenic and benign incident rates for (**C**) SLiMs or (**D**) phosphorylation sites by region class were calculated as the number of missense mutations of a given type (pathogenic or benign) found in a SLiM or at a phosphorylation site in a region class, divided by the total number of mutations of a type and then normalized to the reference human proteome composition by region class for SLiMs or phosphorylation sites.