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 many, but not all, show an enrichment of phase separation in the proteins associated with them. Strikingly, the pathogenic mutation rate 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). Disease-linked mutations, however, are more prevalent in the folded regions (9). The functional impact of mutations in folded regions often can be understood by their perturbations to structure and stability (10–12). For IDRs, disease-linked mutations have been found to disrupt protein-protein interactions (13, 14), change the sequence's propensity for disorder (9), alter the dimensions (e.g., compaction) of the disordered conformational ensemble (15), and modify sites of post-translational modifications (14) or short linear motifs (SLiMs) (16).

IDRs have been associated with many key biological processes, including signal transduction (17, 18), chromatin remodeling (19), cell cycle control (20), and directing the curvature of membrane surfaces (21). IDRs also are essential mediators of liquid-liquid phase separation (22, 23), a phenomenon in which condensed liquid droplets separate from and subsequently coexist within bulk cytosol (24). Biological condensates such as nucleoli, Cajal bodies and stress granules are critical for intracellular compartmentalization and functional specificity (24–26). Human genetic variations that mutate IDR sequences thus potentially alter the function and properties of biological condensates, and possibly contribute to disease phenotypes (27).

Several studies provide evidence that dysregulation of protein phase separation has a role in human disease (28–30). 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 (31). Fusion oncoproteins, e.g., EWS-FLI1 (32, 33), spontaneously form condensates inside cells that promote cancer development (34, 35). Many proteins known to phase separate also have been found in neurodegenerative inclusions (30), for example α-synuclein in Parkinson's disease (36), tau in Alzheimer's disease (37), and TDP-43 (38) or FUS (39) in amyotrophic lateral sclerosis. Additionally, several predictors of phase separation have been used to identify correlations between phase separation and disease (40, 41), including cancer and autism spectrum disorder (27).

Previously, we developed an algorithm, ParSe, that accurately identifies from the protein primary sequence those IDRs likely to exhibit physiological phase separation behavior (42, 43). 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 (43). ParSe shows similar accuracy to other methods that have been developed to predict which protein sequences drive phase separation (43–49). One advantage of the ParSe algorithm is its speed, which enables it to identify drivers of phase separation at proteome scale (50).

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 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 (46, 51–54). However, the types of substitutions that tended to be pathogenic in predicted PS IDRs also tended to be pathogenic in the other protein regions as well. 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 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 (55) 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. Human proteins (1,248 sequences) with entries in DisProt (56) were used to represent IDPs in general. Sequence sets representing phase separation exhibited varying levels of curation: human proteins (3,917 sequences) annotated as "condensate" in CD-CODE (57), human proteins (129 sequences) annotated as "condensate driver" in CD-CODE, and *in vitro* confirmed phase-separating human proteins (66 sequences) assembled previously (42) from Vernon et al (46), DisProt (56), and PhaSePro (58).

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 (55). 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 (59) and ClinVar databases (4) and include human variations from the 1,000 Genomes Project (3, 60). 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 (61). 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 (62).

Sequence-based calculation of phase separation potential

Dysregulation of protein phase separation has been associated with several human diseases (27–30). To explore this relationship broadly, we used the ParSe algorithm (43) to examine how different types of mutations (pathogenic, benign, or of unknown clinical significance) correlate with phase separation. Additionally, we compared the predicted prevalence of phase separation of disease-associated proteins 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 (43). As diagramed in Figure S1, 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 multiple human proteins confirmed to phase separate *in vitro* (51, 63–68).

ParSe also can 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 (Figure S1). The classifier distance was developed to assess confidence in the window assignments (43). 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 a set of 43 confirmed to exhibit homotypic phase separation behavior, curated by Vernon et al (46), had a computed PS potential of at least 100 (Figure S2). We use that minimum potential as a cut-off throughout this paper, identifying PS predicted proteins as those with a PS potential above 100.

Mutations in IDRs link pathogenicity and phase separation

We sought to compare rates of missense mutations in PS IDRs and nonPS IDRs (Figure 1). Strikingly, we found significant enrichment of pathogenic mutations within PS IDRs as compared to nonPS IDRs. Normalized mutation rates were calculated as the proportion of mutations divided by the proportion of residues in a region class (Tables S1 and S2). Figure 1B shows that the normalized rate for pathogenic missense mutations in predicted PS IDRs (0.66 ± 0.02, rate ± standard error) was approximately half the rate found in predicted folded regions (1.22 ± 0.002). The PS IDR rate was almost 3-fold higher than the rate calculated in predicted nonPS IDRs (0.24 ± 0.009). Thus, though the normalized 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 pathogenic versus benign odds ratio also revealed that folded regions exhibited a 2-fold higher pathogenicity compared to PS IDRs, while PS IDRs had a 3-fold higher pathogenicity than nonPS IDRs (Figure S3A).

This ranking of pathogenic mutation rates by region class (folded > PS ID > nonPS ID) was observed in subset proteomes of proteins with confirmed mutations, confirmed IDRs, and confirmed phase separation behavior (Figure 1B). The sets curated for phase separation showed increases in fold enrichment for PS compared to the sets curated for ID and mutation (Figure 1D). Fold enrichment was calculated as the fractional number of PS predicted proteins in a set divided by its fractional number in the reference human proteome. The ratio of pathogenic mutation rates, PS IDR rate divided by nonPS IDR rate, was >1 for each sequence set and correlated weakly with fold enrichment for PS ($r = 0.17$), suggesting that bias from the computational predictor, or that could arise from some proteins being more researched than others, was minimal. Moreover, the probability of finding a pathogenic missense mutation in a sequence set (mutations found divided by total residues) increased for sets curated for confirmed phase separation behavior (Figure 1E). This shows that mutations in PS proteins are more pathogenic.

In contrast, there was no significant difference in the normalized 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, for each of the full reference and subset proteomes (Figure 1C). Figures 1F-L show the reported and predicted domain-level organization of representative proteins *in vitro* verified to phase separate (51, 63–68), as well as the sites of their confirmed pathogenic and benign missense mutations. Qualitatively similar results were obtained with the natural variants index, whereby the normalized 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 S4).

We then sought to determine if these domain-specific differences in pathogenic mutation rates could be generalized to proteins containing significant PS regions and found that all mutations are overrepresented in PS predicted proteins as compared to the human proteome. Using a cutoff of PS potential of 100, ~16% of the human proteome and ~21% of the proteins with pathogenic mutations, a 1.3-fold enrichment, are predicted to exhibit homotypic phase separation behavior (Figure S5). The distribution of PS potential values in the two sets showed a statistically significant difference by the nonparametric Mann-Whitney *U*-test (69), giving a one-tail *p*-value < 2.2e-16. Similar results were obtained for proteins with benign or of uncertain clinical significance mutations in the humsavar index, showing 1.2-fold and 1.6-fold enrichment, respectively. Qualitatively similar results were obtained with the natural variants index, whereby proteins harboring single amino acid missense mutations were enriched in predicted phase separation behavior (Figure S5). Thus, proteins with sequence features predicting phase separation were more likely to have identified mutations. This may be due to the differences in overall size and composition, as nonPS proteins (predicted PS potential below 100) were, on average, shorter in length and more folded as compared to PS predicted proteins (Figure S6).

Mutations in SLiMs within PS IDRs exhibit exceptionally high pathogenicity

SLiMs are conserved recognition motifs involved in signal transduction (70) and found primarily in the IDRs of the human proteome (71). As protein-protein interactions are fundamental to the functional outcomes of phase separation, enabling the formation, regulation, and biological

specificity of biomolecular condensates (26, 72), we next asked if 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; a 2.3-fold enrichment relative to the reference human proteome, where only 16.4% of human proteins are predicted to exhibit phase separation behavior (Figure S5).

Among IDRs, confirmed SLiMs were enriched more so in nonPS than PS IDRs. Here, the normalized rates were calculated as the proportion of confirmed SLiMs divided by the proportion of residues found in a region class. The computed rates in PS ID, nonPS ID, and folded regions were 1.8 ± 0.11, 2.5 ± 0.10, and 0.42 ± 0.01, respectively (Figure 2A).

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 rate for pathogenic missense mutations in SLiMs was 3.03 ± 0.25. The normalized rate for confirmed benign missense mutations in a confirmed SLiM was 1.35 ± 0.15 (Table S2). This result matches the previous finding that, in IDRs, pathogenic missense mutations are more likely to occur within SLiMs than benign missense mutations (16).

Pathogenic mutations were particularly enriched in SLiMs in PS IDRs, though the overall number of such mutations was small. The normalized rate of pathogenic missense mutations in SLiMs in PS IDRs was 10.9 ± 1.6, the highest of all pathogenic 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 2B). For comparison, the normalized rate was 1.6 ± 0.6 for benign missense mutations found in both a confirmed SLiM and a predicted PS IDR. The pathogenic versus benign odds ratio for missense mutations in a SLiM in a predicted PS IDR was 5.6 ± 2.2 (Figure S3B), reflecting the significantly higher normalized pathogenic rate. The normalized rate for pathogenic mutation in both a SLiM and predicted nonPS IDR, and a SLiM and predicted folded region was 1.7 ± 0.4 and 3.1 ± 0.4, respectively (Figure 2B). Thus, pathogenicity increased for SLiM positions in each region class (Figure S7) relative to the pathogenic mutation rate that included nonSLiM positions (Figure 1A). However, pathogenicity was highest, and by a significant margin, for missense mutations in a SLiM that also occurred in a predicted PS IDR (Figure 2B). As such, even though SLiMs are more likely to be found in nonPS IDRs than PS IDRs, their mutation was more likely pathogenic in PS IDRs.

Notably, the computed pathogenic mutation rate drops to the base PS IDR rate (~0.6 ± 0.1) with increasing residue distance from the SLiM (Figure 2B inset). Thus, the exceptionally high pathogenic missense mutation rate was concentrated on the SLiM in the IDR and not the nonSLiM positions. As conserved recognition sites, SLiMs and other protein-protein interaction interfaces facilitate biological phase separation (73) through various key roles, e.g., condensate formation, stability, internal structure, and composition, which may explain their exceptionally high mutational pathogenicity.

Mutations at phosphorylation sites within PS IDRs are mostly benign

Protein phosphorylation is used biologically as a mechanism to control many critical life processes (74, 75) and phosphorylation often occurs within protein regions that are ID (76). Phosphorylation has been found to both promote (77, 78) and inhibit (79, 80) phase separation by modulating protein charge, flexibility, and oligomerization state.

We found that 18.8% of human proteins with confirmed phosphorylation sites were predicted by ParSe to exhibit phase separation behavior; a 1.1-fold enrichment relative to the reference human proteome (Figure S5). The normalized rate for confirmed phosphorylation sites found in predicted PS IDRs, nonPS IDRs, and folded regions was 2.73 ± 0.07, 1.77 ± 0.07, and 0.67 ± 0.03, respectively (Figure 2A). Thus, 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 ± 0.03 (Figure 2C). The normalized rate for pathogenic missense mutations at a confirmed phosphorylation site and in a predicted nonPS IDR was 0.23 ± 0.03, and in a predicted folded region was 0.99 ± 0.04. The pathogenic versus benign odds ratio was similarly low, 0.26 ± 0.04 and 0.23 ± 0.03, for phosphorylation sites in PS and nonPS IDRs, respectively, when compared to phosphorylation sites in folded regions, 1.0 ± 0.05 (Figure S3C). 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. Possibly explaining this result, phosphorylation in folded domains has been observed to occur at buried positions that become transiently accessible, which can directly affect protein structure, dynamics, and biological activity (81). In contrast, phosphorylation in IDRs is often characterized by multisite phosphorylation (82–85) and, accordingly, mutational perturbation of a single site might be less deleterious to function.

Disease-associated proteins are enriched for phase separation

Tsang et al, using the PScore algorithm (46), found that disease-associated human proteins are enriched in phase separation potential (27). Similarly, ParSe predicts that ~23% of human proteins annotated as disease-associated in UniProt are likely to exhibit homotypic phase separation behavior, which is a 1.4-fold enrichment compared to the reference human proteome (Figure S5).

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 S8). This alternative approach was used because, first, when comparing sequence sets, it eliminates the need for a cutoff PS potential for defining proteins as PS or nonPS. The computed AUC is proportional to the fold enrichment for PS whether the cutoff is 25, 50, 100, etc. (Figure S9). AUC > 0.5 indicates a sequence set enriched in computed phase separation potential relative to the reference human proteome. Secondly, the computed AUC can be compared to the mean AUC from sets of random human proteins, facilitating comparisons between sequence sets of vastly different sizes (Figure 3A). AUC values more than two standard deviations from the mean AUC of random sets

were statistically equivalent to *p*-values < 0.05 (Figure 3A inset). Thus, AUC values, when compared to the random expectation, can identify when sets of PS potentials have statistically significant differences.

This approach was used to analyze the sets of disease-associated human proteins (Table S3). When sub-grouped by disease type, 79 of 100 disease types produced AUC > 0.5 (Figure S10), 51 produced AUC more than a standard deviation above the mean AUC from random proteins (Figure 3A), and 32 produced AUC more than 2 standard deviations above the mean random AUC (Figure 3A inset). In particular, this result recapitulates the previous finding (27) that proteins associated with cancer and autism spectrum disorder have an over-representation of predicted phase-separating behavior (Figure 3B and Table S3).

We used the standard deviation to identify disease types with unusual and extreme enrichment for PS as compared to random. This was done to avoid sub-groups with enrichment for PS (> 1-fold) but also *p*-values > 0.05, indicating data not sufficient to reject the null hypothesis. Several disease sub-groups had significantly elevated AUC values that were more than three standard deviations above the mean of matched random protein sets. These included neurodevelopmental disorders (such as intellectual disability, autism spectrum disorder, epilepsy, microcephaly, and holoprosencephaly), cancer (proto-oncogenes and tumor suppressors), and connective tissue diseases (such as Ehlers-Danlos syndrome) (Figure 3B and Table S3). Many of the proteins associated with cancers, neurodegeneration, and neurodevelopment disorders are involved in synaptic function (86), transcription regulation (87), and chromatin remodeling (88, 89); biological processes that have been linked to phase separation (90–94). Also, phase separation is thought to drive skin barrier formation (95). Thus, aberrant phase separation of skin proteins potentially has a role in skin disease (epidermolysis bullosa, palmoplantar keratoderma, Ehlers-Danlos syndrome).

In comparison, disease sub-groups with computed AUC three standard deviations (or more) below 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. In addition to lacking predicted PS IDRs, the proteins in these three sub-groups were on average shorter and more folded, ~370 residues and ~85% folded, compared to proteins in the reference proteome, ~560 residues and ~75% folded, and especially when compared to human proteins *in vitro* confirmed for phase separation, ~680 residues and ~50% folded.

Sequence dependence to pathogenicity does not depend on region identity

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 that the types of mutations that tended to be pathogenic in predicted PS IDRs also tended to be pathogenic in the other protein regions as well.

We first considered the entire set of mutations in the humsavar index (32,665 pathogenic mutations from 3,345 proteins) and found that the most frequent pathogenic substitutions primarily involved charged residues (R→{C, H, Q, W}, G→R, and E→K) as well as proline (P→L and L→P) (Figure

S11A). Many of these same substitutions (e.g., R→H, R→Q, P→L, E→K) also were found in mutations of benign clinical significance (39,656 benign mutations from 11,682 proteins), emphasizing the importance of sequence context (Figure S11B). Figure 4A shows the difference in the calculated substitution frequencies, pathogenic minus benign, for all combinations of substitution types. The highest net positive values (i.e., primarily pathogenic) were the nonconservative changes of L→P, G→R, and R→W, while he lowest net negative values (i.e., primarily benign) were substitutions with generally minimal changes in chemical character (V→I, I→V, A→T, and T→A).

We next tested if pathogenic and benign substitution trends differed by region identity, to explore why the same substitution sometimes was harmless and sometimes harmful. We found that single amino acid missense mutations in regions identified by ParSe as matching the PS ID class reported different substitution preferences (Figure S12). 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 (96, 97). 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 the most net pathogenic changes involved arginine (G→R, R→C, and C→R), while the most net benign involved the small polar hydroxyl-containing residues (T→A, S→A, S→N, and P→S). We used the Mantel test (98) to statistically compare the matrices of net pathogenic and net benign substitutions, for all positions versus PS IDR positions (Figures 4A and 4B), and found a strong, positive correlation coefficient of 0.87 (p-value = 0.001), indicating that the patterns of values were highly similar.

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. 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 (52–54). 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 in regions matching the nonPS ID and folded classes (Figure S13), we again found strong, positive correlations, 0.74 (p-value = 0.001) and 0.78 (p-value = 0.001), respectively, in the patterns of substitutions that were net pathogenic and net benign when compared to missense

mutations found in PS IDRs. This shows that despite robust compositional differences between the three classes of protein regions, PS ID, nonPS ID, and folded (43), the types of mutations that were pathogenic and the types that were benign did not exhibit significant variations by region class. Generally, substitutions involving Arg, Cys, Gly, Tyr, Trp, and Asp were pathogenic, while Ala, Ser, Thr, Val, and Ile were benign.

Predicting pathogenicity of missense variations by protein region class

We next asked if the pathogenic minus benign difference frequencies could be used to predict clinical significance from the substitution. While more sophisticated predictors of pathogenicity have been developed, typically focused on folded domains (99) or evolutionary models (100, 101), 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.

First, considering all protein regions, we used the difference frequency data in Figure 4A and found 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 the 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 S14).

Focusing on the PS ID class, the combined true positive and true negative rate of missense substitutions 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 regions for single amino acid missense mutations (Figure 1A), 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. This result is consistent with our finding that, in general, the net pathogenic and net benign substitution types did not change significantly with region identity.

The 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% 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.

Discussion

It has been proposed that mutations that alter IDR-induced phase separation may have a higher likelihood of causing disease (27–30). Indeed, the results from several studies indicate a role for aberrant physiological protein phase separation in the progression of some human disorders (31–39). 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 many 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, cancer, neurodegeneration, and the skin (Figure 3B). 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 the longer average length and higher ID content of predicted PS proteins. Pathogenic missense mutation rates were highest in the folded regions of proteins (Figure 1B), matching prior reports (9). Benign missense mutation rates did not depend on region identity.

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 (Figure 1B). Thus, although IDRs, and even PS IDRs, show low evolutionary conservation of sequence, mutations in these regions can be functionally significant. Pathogenicity was highest for missense mutations in a predicted PS IDR when also in a confirmed SLiM, and significantly higher than for missense mutations in a SLiM and either of the other region classes (Figure 2B). Though highest in PS IDRs, pathogenicity increased for mutations at SLiM positions in each region class (Figure S7), possibly reflecting the conserved sequence patterning of SLiMs (102) and their crucial roles in a wide variety of cellular processes (73).

The significantly higher pathogenicity of SLiMs located in PS IDRs compared to those in folded regions may reflect structural accessibility constraints. SLiMs function as short, modular interaction motifs that typically rely on being solvent-exposed to engage with their recognition partners (103). IDRs offer a flexible and accessible environment where SLiMs can participate in transient interactions. SLiMs embedded within folded domains may be conformationally buried and subsequently less accessible for interaction. This suggests that SLiMs may necessitate disorder for flexibility and functionality, which may explain the increased likelihood of pathogenicity when SLiMs are located within PS IDRs.

The finding that pathogenic mutations were not enriched at phosphorylation sites in PS IDRs (Figure 2C) contrasts with the observed enrichment of pathogenic mutations in SLiMs in the same regions. Phosphorylation in IDRs often involves multisite modification rather than dependence on a single site (82–85). In proteins such as tau, hyperphosphorylation across many residues regulates phase separation behavior (79). In contrast, SLiMs generally function as discrete, site-specific modules for molecular recognition, therefore a mutation in a single SLiM can disrupt binding

events and lead to more pronounced pathogenic effects. These structural and functional differences between the two sequence elements may help explain why they differ in their susceptibility to pathogenic outcomes following mutation.

The clinical significance of missense mutations in predicted PS IDRs revealed a dependence on amino acid type that largely mirrored the dependence observed in nonPS and folded regions (Figures 4 and S13). 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 (104). Numerous mutagenesis studies find Arg, Tyr, and Phe as mediators of droplet formation (46, 51–54). Mutating charge patterns through Asp substitution has been found to modulate phase separation potential (51). Protein-protein interactions involving His have been shown to contribute to phase separation behavior (105), especially toward pH responsive self-assembling processes (106). 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. 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. 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 (55), 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 (61), in January 2024. Human proteins with confirmed phosphorylation sites were obtained from the Eukaryotic Phosphorylation Site Database 2.0 (62) in March 2024. Human proteins with confirmed IDRs were found by search of the DisProt database (56), in June 2025. Human proteins with confirmed phase separation behavior were found by search of the CD-CODE database (57), in June 2025.

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 elsewhere (43) and available as a webtool (50) at stevewhitten.github.io/Parse_v2_web. 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.

Mann-Whitney *U*-test

One-tail *p*-values were calculated using the Mann-Whitney *U*-test (69) as implemented in the R statistical computing and graphics program (107).

Mantel test

Mantel statistics (98) for comparing matrices were calculated by Pearson's product-moment correlation and with 999 permutations. These calculations used the vegan package (108) as implemented in the R statistical computing and graphics program (107).

Fold Enrichment

Fold Enrichment for a protein set was calculated as the fractional number of PS predicted proteins in a set divided by the fractional number of PS predicted proteins in the reference proteome. Protein type, PS or nonPS, was determined by the ParSe algorithm, version 2, where PS proteins had PS potential ≥100.

Calculation of Normalized Missense Mutation Rate

The rate represents a proportion, *p*, of mutations found in a region,

$$p = m_{region}/m, \qquad [1]$$

where m is the total number of mutations, m_{region} is the subset of m found in a region, and regions can be PS IDRs, nonPS IDRs, folded regions, SLiMs in PS IDRs, etc. Pathogenic rates were calculated using missense mutations labeled likely pathogenic or pathogenic; benign rates used missense mutations labeled likely benign or benign. Next, p was multiplied by a factor, k, that normalized rates according to the proportion of residues found in a region. Specifically,

$$p_{normalized} = k \cdot p = \left(r/r_{region}\right) \cdot p, \qquad [2]$$

where r is the total number of residues in the sequence set (e.g., the reference proteome), and r_{region} is the subset of r found in a region type. The standard error of the proportion (non-scaled) is,

$$SE = \sqrt{p \cdot (1 - p)/m}, \quad [3]$$

The adjusted standard error for normalized p is,

$$SE_{normalized} = k \cdot SE. \quad [4]$$

Calculation of SLiM and Phosphorylation Rate

Rates of SLiMs and phosphorylation by region were calculated in a manner identical to the mutation rate. For SLiMs, equation 1 was modified as,

$$p = s_{region}/s, \quad [5]$$

where s is the total number of SLiMs and s_{region} the subset found in a region, and for phosphorylation as,

$$p = phos_{region}/phos, \quad [6]$$

where $phos$ is the total number of phosphorylation sites and $phos_{region}$ the subset found in a region. Each rate was normalized by k using equation 2. For SLiMs, SE was calculated using equation 3 with s substituted for m. For phosphorylation, SE was calculated with $phos$ substituted for m. The adjusted standard error was calculated as given by equation 4.

Calculation of Pathogenic versus Benign Odds Ratio

Odds ratios were used to compare the probabilities of pathogenic versus benign mutations. Probabilities were determined by,

$$P = m_{region}/r_{region}, \quad [5]$$

were m_{region} and r_{region} are defined above. Pathogenic probabilities were calculated using missense mutations labeled likely pathogenic or pathogenic; benign probabilities used missense mutations labeled likely benign or benign. The odds ratio, OR, was determined by,

$$OR = \left(P_{pathogenic}/\left(1 - P_{pathogenic}\right)\right)/\left(P_{benign}/\left(1 - P_{benign}\right)\right), \quad [6]$$

The standard error for the natural log of OR is (109),

$$SE_LOR = \sqrt{\frac{1}{m_{region|pathogenic}} + \frac{1}{(r_{region} - m_{region|pathogenic})} + \frac{1}{m_{region|benign}} + \frac{1}{(r_{region} - m_{region|benign})}}, \quad [7]$$

where $m_{region|pathogenic}$ and $m_{region|benign}$ represent the number of pathogenic and benign mutations found in a region, respectively. The standard error for the OR was estimated by (110),

$$SE_OR \cong OR \cdot SE_LOR. \quad [8]$$

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,

$$Freq = m_{x-to-y}/m. \quad [9]$$

Substitution frequencies by region class, for PS ID, nonPS 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,

$$Freq_{region} = m_{x-to-y|region}/m_{region}. \quad [10]$$

Protein region class by position in a primary sequence was determined using ParSe version 2.

Data Availability

Source code for both the ParSe v2 algorithm and its webtool applications can be downloaded from github.com/stevewhitten.

Additional Information

Supporting Information – This article contains supporting information (43, 55, 61, 62, 69, 111, 112).

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. Vacic, V., Markwick, P. R. L., Oldfield, C. J., Zhao, X., Haynes, C., Uversky, V. N., and Iakoucheva, L. M. (2012) Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. *PLoS Comput Biol*. **8**, e1002709
10. Yue, P., Li, Z., and Moult, J. (2005) Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol*. **353**, 459–473
11. Stein, A., Fowler, D. M., Hartmann-Petersen, R., and Lindorff-Larsen, K. (2019) Biophysical and Mechanistic models for disease-causing protein variants. *Trends Biochem Sci*. **44**, 575–588
12. 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
13. Wong, E. T. C., So, V., Guron, M., Kuechler, E. R., Malhis, N., Bui, J. M., and Gsponer, J. (2020) Protein-Protein Interactions Mediated by Intrinsically Disordered Protein Regions Are Enriched in Missense Mutations. *Biomolecules*. **10**, 1097

14. Ahmed, S. S., Rifat, Z. T., Lohia, R., Campbell, A. J., Dunker, A. K., Rahman, M. S., and Iqbal, S. (2022) Characterization of intrinsically disordered regions in proteins informed by human genetic diversity. *PLoS Comput Biol*. **18**, e1009911
15. Flores, E., Acharya, N., Castañeda, C. A., and Sukenik, S. (2025) Single-point mutations in disordered proteins: Linking sequence, ensemble, and function. *Curr Opin Struct Biol*. **91**, 102987
16. Uyar, B., Weatheritt, R. J., Dinkel, H., Davey, N. E., and Gibson, T. J. (2014) Proteome-wide analysis of human disease mutations in short linear motifs: neglected players in cancer? †Electronic supplementary information (ESI) available: Supplementary files 1–22 and supplementary Fig. 1–3. See DOI: 10.1039/c4mb00290c Click here for additional data file. Click here for additional data file. Click here for additional data file. *Mol Biosyst*. **10**, 2626–2642
17. 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
18. 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
19. 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
20. 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
21. 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
22. 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
23. Gomes, E., and Shorter, J. (2019) The molecular language of membraneless organelles. *J. Biol. Chem.* **294**, 7115–7127
24. 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
25. 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
26. Mitrea, D. M., and Kriwacki, R. W. (2016) Phase separation in biology; functional organization of a higher order. *Cell Communication and Signaling*. **14**, 1
27. 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
28. Alberti, S., and Dormann, D. (2019) Liquid–Liquid Phase Separation in Disease. *Annual Review of Genetics*. **53**, 171–194
29. 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

30. 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
31. 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
32. 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
33. 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
34. 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
35. 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
36. 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
37. 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
38. 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
39. 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
40. 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
41. 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

42. 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
43. 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
44. 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
45. 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
46. 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
47. 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
48. 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
49. 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
50. 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
51. 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
52. 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
53. 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
54. 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
55. The UniProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*. **51**, D523–D531
56. Hatos, A., Hajdu-Soltész, B., Monzon, A. M., Palopoli, N., Álvarez, L., Aykac-Fas, B., Bassot, C., Benítez, G. I., Bevilacqua, M., Chasapi, A., Chemes, L., Davey, N. E., Davidović, R., Dunker, A. K., Elofsson, A., Gobeill, J., Foutel, N. S. G., Sudha, G., Guharoy, M., Horvath, T., Iglesias, V., Kajava, A. V., Kovacs, O. P., Lamb, J., Lambrughi, M., Lazar, T., Leclercq, J. Y., Leonardi, E., Macedo-Ribeiro, S., Macossay-Castillo, M.,

Maiani, E., Manso, J. A., Marino-Buslje, C., Martínez-Pérez, E., Mészáros, B., Mičetić, I., Minervini, G., Murvai, N., Necci, M., Ouzounis, C. A., Pajkos, M., Paladin, L., Pancsa, R., Papaleo, E., Parisi, G., Pasche, E., Barbosa Pereira, P. J., Promponas, V. J., Pujols, J., Quaglia, F., Ruch, P., Salvatore, M., Schad, E., Szabo, B., Szaniszló, T., Tamana, S., Tantos, A., Veljkovic, N., Ventura, S., Vranken, W., Dosztányi, Z., Tompa, P., Tosatto, S. C. E., and Piovesan, D. (2020) DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* **48**, D269–D276

57. Rostam, N., Ghosh, S., Chow, C. F. W., Hadarovich, A., Landerer, C., Ghosh, R., Moon, H., Hersemann, L., Mitrea, D. M., Klein, I. A., Hyman, A. A., and Toth-Petroczy, A. (2023) CD-CODE: crowdsourcing condensate database and encyclopedia. *Nat Methods*. **20**, 673–676
58. Mészáros, B., Erdős, G., Szabó, B., Schád, É., Tantos, Á., Abukhairan, R., Horváth, T., Murvai, N., Kovács, O. P., Kovács, M., Tosatto, S. C. E., Tompa, P., Dosztányi, Z., and Pancsa, R. (2020) PhaSePro: the database of proteins driving liquid–liquid phase separation. *Nucleic Acids Res*. **48**, D360–D367
59. 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
60. 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

61. 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
62. 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

63. Tsang, B., Arsenault, J., Vernon, R. M., Lin, H., Sonenberg, N., Wang, L.-Y., Bah, A., and Forman-Kay, J. D. (2019) Phosphoregulated FMRP phase separation models activity-dependent translation through bidirectional control of mRNA granule formation. *Proc Natl Acad Sci U S A*. **116**, 4218–4227
64. Wang, L., Kang, J., Lim, L., Wei, Y., and Song, J. (2018) TDP-43 NTD can be induced while CTD is significantly enhanced by ssDNA to undergo liquid-liquid phase separation. *Biochem Biophys Res Commun*. **499**, 189–195
65. Kostylev, M. A., Tuttle, M. D., Lee, S., Klein, L. E., Takahashi, H., Cox, T. O., Gunther, E. C., Zilm, K. W., and Strittmatter, S. M. (2018) Liquid and Hydrogel Phases of PrP^C Linked to Conformation Shifts and Triggered by Alzheimer's Amyloid-β Oligomers. *Mol Cell*. **72**, 426-443.e12
66. Molliex, A., Temirov, J., Lee, J., Coughlin, M., Kanagaraj, A. P., Kim, H. J., Mittag, T., and Taylor, J. P. (2015) Phase Separation by Low Complexity Domains Promotes Stress Granule Assembly and Drives Pathological Fibrillization. *Cell*. **163**, 123–133
67. Sheu-Gruttadauria, J., and MacRae, I. J. (2018) Phase Transitions in the Assembly and Function of Human miRISC. *Cell*. **173**, 946-957.e16
68. Lin, Y.-H., Qiu, D.-C., Chang, W.-H., Yeh, Y.-Q., Jeng, U.-S., Liu, F.-T., and Huang, J.-R. (2017) The intrinsically disordered N-terminal domain of galectin-3 dynamically mediates multisite self-association of the protein through fuzzy interactions. *J Biol Chem*. **292**, 17845–17856
69. Mann, H. B., and Whitney, D. R. (1947) On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*. **18**, 50–60
70. 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
71. 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
72. Li, P., Banjade, S., Cheng, H.-C., Kim, S., Chen, B., Guo, L., Llaguno, M., Hollingsworth, J. V., King, D. S., Banani, S. F., Russo, P. S., Jiang, Q.-X., Nixon, B. T., and Rosen, M. K. (2012) Phase transitions in the assembly of multivalent signalling proteins. *Nature*. **483**, 336–340
73. Van Roey, K., Uyar, B., Weatheritt, R. J., Dinkel, H., Seiler, M., Budd, A., Gibson, T. J., and Davey, N. E. (2014) Short Linear Motifs: Ubiquitous and Functionally Diverse Protein Interaction Modules Directing Cell Regulation. *Chem. Rev.* **114**, 6733–6778
74. 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
75. Whitmarsh, A. J., and Davis, R. J. (2000) Regulation of transcription factor function by phosphorylation. *Cell Mol Life Sci*. **57**, 1172–1183
76. 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
77. Carlson, C. R., Asfaha, J. B., Ghent, C. M., Howard, C. J., Hartooni, N., Safari, M., Frankel, A. D., and Morgan, D. O. (2020) Phosphoregulation of Phase Separation by the

SARS-CoV-2 N Protein Suggests a Biophysical Basis for its Dual Functions. *Molecular Cell*. **80**, 1092-1103.e4
78. Liu, D., Lum, K. K., Treen, N., Núñez, C. T., Yang, J., Howard, T. R., Levine, M., and Cristea, I. M. (2023) IFI16 phase separation via multi-phosphorylation drives innate immune signaling. *Nucleic Acids Res*. **51**, 6819–6840
79. Abasi, L. S., Elathram, N., Movva, M., Deep, A., Corbett, K. D., and Debelouchina, G. T. (2024) Phosphorylation regulates tau's phase separation behavior and interactions with chromatin. *Commun Biol*. **7**, 251
80. Monahan, Z., Ryan, V. H., Janke, A. M., Burke, K. A., Rhoads, S. N., Zerze, G. H., O'Meally, R., Dignon, G. L., Conicella, A. E., Zheng, W., Best, R. B., Cole, R. N., Mittal, J., Shewmaker, F., and Fawzi, N. L. (2017) Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity. *The EMBO Journal*. **36**, 2951–2967
81. Henriques, J., and Lindorff-Larsen, K. (2020) Protein Dynamics Enables Phosphorylation of Buried Residues in Cdk2/Cyclin-A-Bound p27. *Biophys J*. **119**, 2010–2018
82. Wang, J.-Z., Xia, Y.-Y., Grundke-Iqbal, I., and Iqbal, K. (2013) Abnormal hyperphosphorylation of tau: sites, regulation, and molecular mechanism of neurofibrillary degeneration. *J Alzheimers Dis*. **33 Suppl 1**, S123-139
83. Gruijs da Silva, L. A., Simonetti, F., Hutten, S., Riemenschneider, H., Sternburg, E. L., Pietrek, L. M., Gebel, J., Dötsch, V., Edbauer, D., Hummer, G., Stelzl, L. S., and Dormann, D. (2022) Disease-linked TDP-43 hyperphosphorylation suppresses TDP-43 condensation and aggregation. *EMBO J*. **41**, e108443
84. Lu, H., Yu, D., Hansen, A. S., Ganguly, S., Liu, R., Heckert, A., Darzacq, X., and Zhou, Q. (2018) Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. *Nature*. **558**, 318–323
85. Kandangwa, M., and Liu, Q. (2019) HCV NS5A hyperphosphorylation is involved in viral translation modulation. *Biochem Biophys Res Commun*. **520**, 192–197
86. Taoufik, E., Kouroupi, G., Zygogianni, O., and Matsas, R. (2018) Synaptic dysfunction in neurodegenerative and neurodevelopmental diseases: an overview of induced pluripotent stem-cell-based disease models. *Open Biology*. **8**, 180138
87. Casamassimi, A., Ciccodicola, A., and Rienzo, M. (2023) Transcriptional Regulation and Its Misregulation in Human Diseases. *Int J Mol Sci*. **24**, 8640
88. Mossink, B., Negwer, M., Schubert, D., and Nadif Kasri, N. (2021) The emerging role of chromatin remodelers in neurodevelopmental disorders: a developmental perspective. *Cell Mol Life Sci*. **78**, 2517–2563
89. Nair, S. S., and Kumar, R. (2012) Chromatin remodeling in Cancer: A Gateway to regulate gene Transcription. *Mol Oncol*. **6**, 611–619
90. Chen, X., Wu, X., Wu, H., and Zhang, M. (2020) Phase separation at the synapse. *Nat Neurosci*. **23**, 301–310
91. Wagh, K., Garcia, D. A., and Upadhyaya, A. (2021) Phase separation in transcription factor dynamics and chromatin organization. *Current Opinion in Structural Biology*. **71**, 148–155
92. Stortz, M., Presman, D. M., and Levi, V. (2024) Transcriptional condensates: a blessing or a curse for gene regulation? *Commun Biol*. **7**, 187
93. Conte, M., Irani, E., Chiariello, A. M., Abraham, A., Bianco, S., Esposito, A., and Nicodemi, M. (2022) Loop-extrusion and polymer phase-separation can co-exist at the single-molecule level to shape chromatin folding. *Nat Commun*. **13**, 4070

94. Rippe, K. (2022) Liquid–Liquid Phase Separation in Chromatin. *Cold Spring Harb Perspect Biol*. **14**, a040683
95. Quiroz, F. G., Fiore, V. F., Levorse, J., Polak, L., Wong, E., Pasolli, H. A., and Fuchs, E. (2020) Liquid-liquid phase separation drives skin barrier formation. *Science*. 10.1126/science.aax9554
96. 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
97. 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
98. Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Res*. **27**, 209–220
99. 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
100. Reva, B., Antipin, Y., and Sander, C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res*. **39**, e118
101. 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
102. Ren, S., Yang, G., He, Y., Wang, Y., Li, Y., and Chen, Z. (2008) The conservation pattern of short linear motifs is highly correlated with the function of interacting protein domains. *BMC Genomics*. **9**, 452
103. Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., and Gibson, T. J. (2012) Attributes of short linear motifs. *Mol Biosyst*. **8**, 268–281
104. 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
105. 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
106. 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
107. R Core Team (2022) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria
108. Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Borman, T., Carvalho, G., Chirico, M., Caceres, M. D., Durand, S., Evangelista, H. B. A., FitzJohn, R., Friendly, M., Furneaux, B., Hannigan, G., Hill, M. O., Lahti, L., Martino, C., McGlinn, D., Ouellette, M.-H., Cunha, E. R., Smith, T., Stier, A., Braak, C. J. F. T., and Weedon, J. (2025) *vegan: Community Ecology Package*, 10.32614/CRAN.package.vegan

109. Savojardo, C., Babbi, G., Martelli, P. L., and Casadio, R. (2019) Functional and Structural Features of Disease-Related Protein Variants. *Int J Mol Sci*. **20**, 1530
110. Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M., and Yu, H. (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat Biotechnol*. **30**, 159–164
111. 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
112. 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

Figures

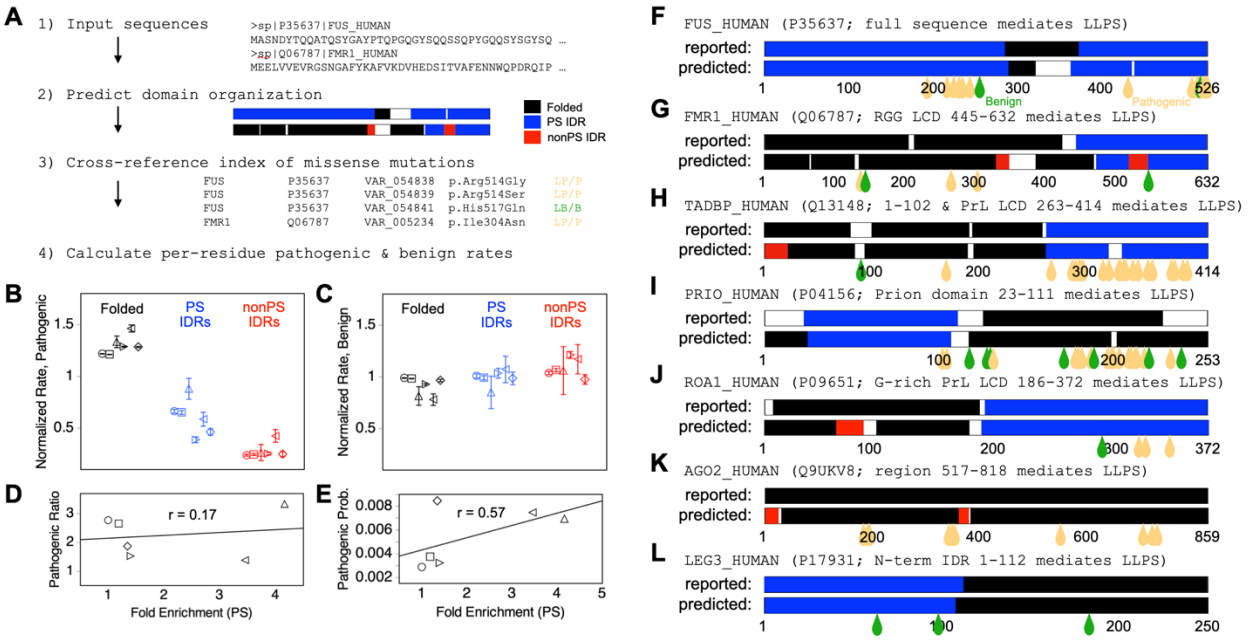

Figure 1. Pathogenic and benign mutation rates by protein region. (**A**) Steps used to calculate mutation rates. Domain organization was predicted from sequence by ParSe. Pathogenic and benign missense mutations were obtained from the humsavar index. Per residue (**B**) pathogenic and (**C**) benign mutation rates by region class were calculated for, from left to right, the reference human proteome (○), proteins with entries in humsavar (□), *in vitro* verified phase-separating human proteins (△), human proteins annotated condensate in CD-CODE (▷), human proteins annotated condensate driver in CD-CODE (◁), and human proteins with entries in DisProt (◇). Error bars show the standard error. (**D**) Sequence sets curated for phase separation had the highest enrichment for proteins predicted PS relative to the reference proteome ("Fold Enrichment (PS)"); the ratio of pathogenic mutation rates ("Pathogenic Ratio"), PS IDR rate divided by nonPS IDR rate, correlated weakly with the fold enrichment (r = 0.17). (**E**) Per residue probability of finding pathogenic mutations exhibited a correlation with the fold enrichment (r = 0.57). Reported versus predicted domain organization for (**F**) FUS, (**G**) FMRP, (**H**) TDP-43, (**I**) PrP, (**J**) hnRNP A1, (**K**) AGO2, and (**L**) Galectin-3. Teardrops show sites of pathogenic (light orange) and benign (green) missense mutations.

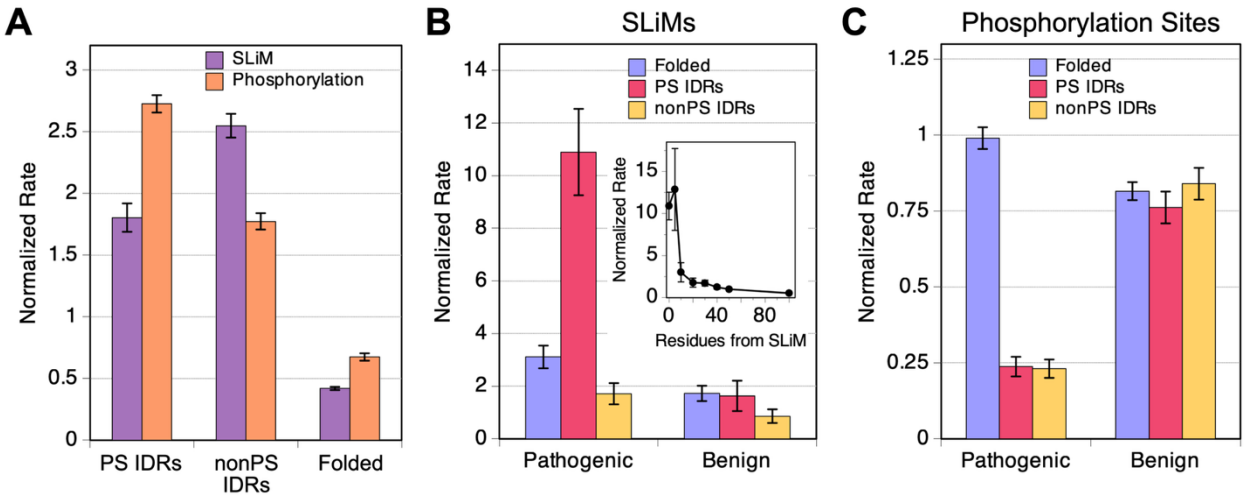

Figure 2. SLiM and phosphorylation rates by protein region and their rates of pathogenic and benign mutation. (**A**) Rates for finding a SLiM (purple) or phosphorylation site (salmon) by protein region class. Pathogenic and benign mutation rates for (**B**) SLiMs or (**C**) phosphorylation sites by protein region class. The panel **B** inset shows the pathogenic mutation rate for residues at positions located immediately outside the SLiM. Rates were calculated using the reference human proteome.

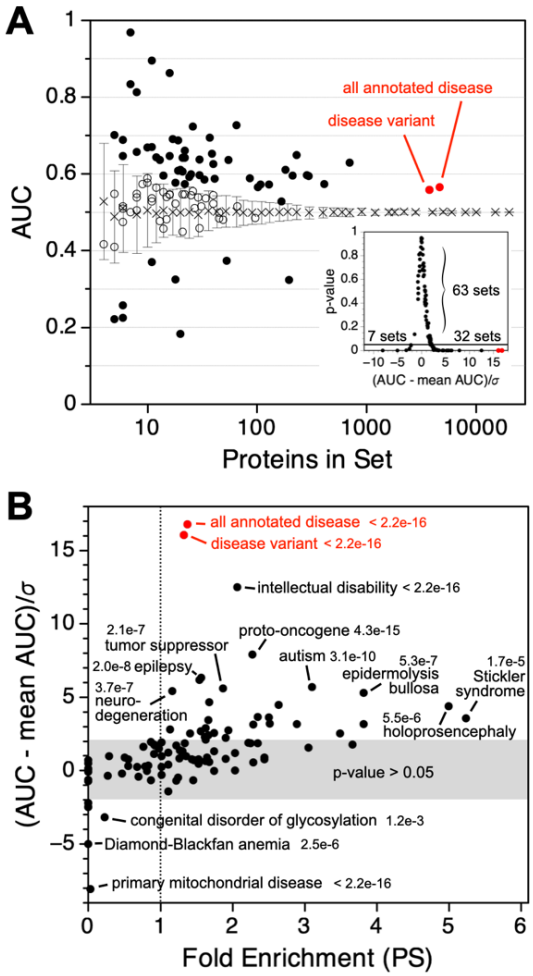

Figure 3. Disease-associated proteins are enriched for phase separation. (**A**) AUC for sets of proteins annotated as disease associated in UniProt were plotted according to the number of proteins in each set (open and filled circles). The mean and standard deviation (x with error bars) from random 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 a set containing proteins for which at least one genetic variant involved in a disease has been reported. The inset shows the relationship between the *p*-value and the number of standard deviations from the mean random AUC. Both AUC and the *p*-value were obtained by comparing the distribution of PS potential in a set against its distribution in the reference human proteome. (**B**) Fold enrichment for PS is plotted according to the number of standard deviations from the mean random AUC. For labeled outliers, the *p*-value is shown.

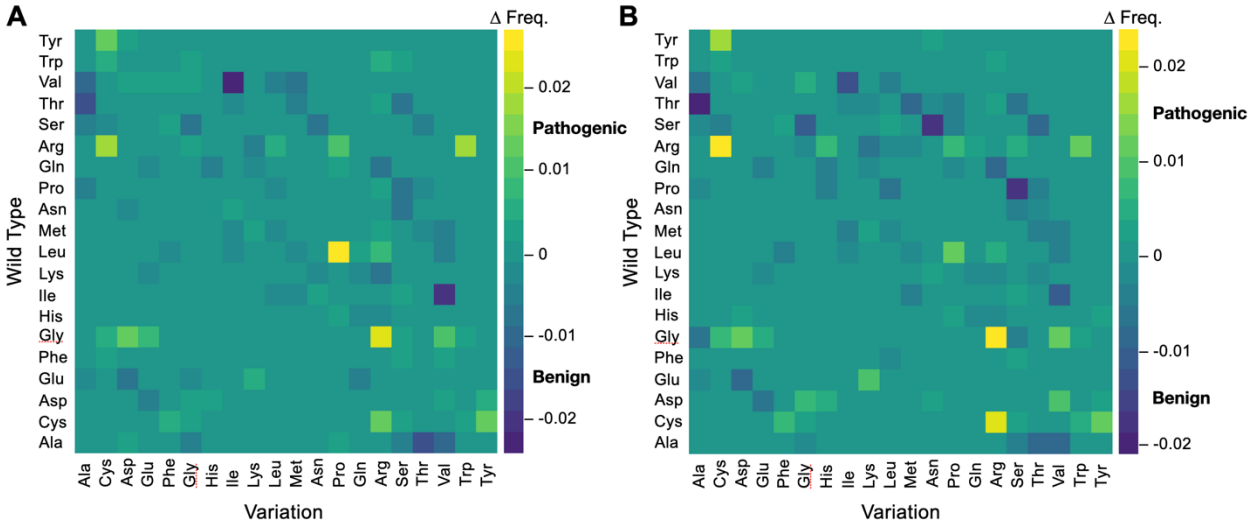

Figure 4. Difference frequencies of pathogenic minus benign single amino acid missense mutations. (**A**) All single amino acid missense mutations in the humsavar index, and (**B**) only those found in a protein region matching the PS ID class and excluding collagen proteins. Difference frequencies (Δ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

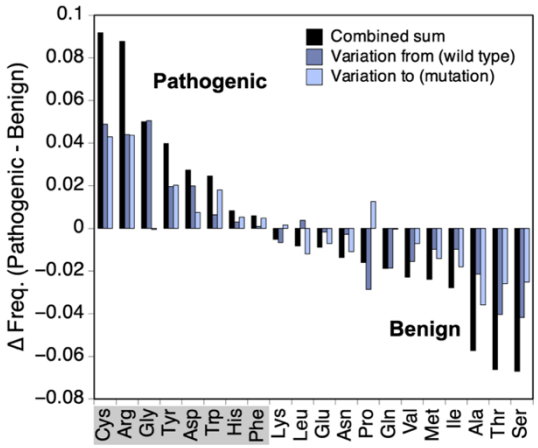

Figure 5. Pathogenicity by amino acid substitution in PS IDRs. Pathogenic minus benign difference frequencies (Δ Freq.) for positions in predicted PS IDRs were summed by wild type amino acid (blue), mutation amino acid (light blue), and the combined sum (black). Contributions from collagen proteins were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

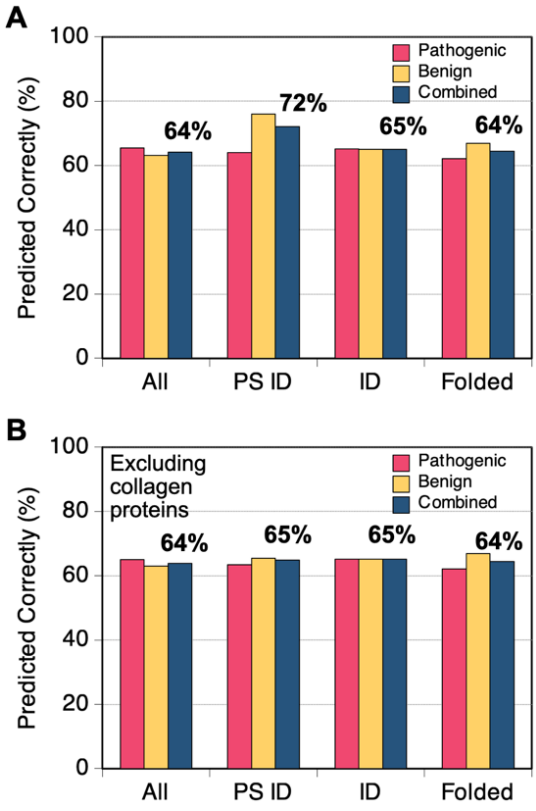

Figure 6. Predicting clinical significance of single amino acid missense mutations. (**A**) Pathogenic minus benign difference frequencies were calculated for all single amino acid missense mutations in the humsavar index and for the subsets residing in protein regions matching the PS ID, ID (i.e., nonPS 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.