Disease associated mutations in intrinsically disordered proteins show evidence of enrichment in hydrophobic blobs

Ruchi Lohia[1], Kaitlin Bassi[1], Matt Hansen[2], Grace Brannigan[1,3*],

1 Center for Computational and Integrative Biology, Rutgers University, Camden, NJ, USA
2 Department of Genetics and Center of Excellence in Environmental Toxicology, University of Pennsylvania
3 Department of Physics, Rutgers University, Camden, NJ, USA

* grace.brannigan@rutgers.edu(GB)

Abstract

Author summary

Introduction

An ambitious task for human genetics is discovering the genetic basis for heritable traits and disease risks. Genome-wide association studies (GWAS) can identify variants associated with a effect regardless of the genomic or protein context of the variant. Associated variants are often not directly causal but instead passively "tag" haplotypes containing the true causal variants that are either unknown or observed but not directly tested for association. The pattern of linkage disequilibrium that defines "tagging" is population dependent. Any information on the protein-level function of the local sequence surrounding an associated variant can help fine-map and rank putative causal variants from a list of associations.

Many bioinformatics approaches apply known principles of protein physical chemistry to the local sequence near the variant, considering effects of the mutation on properties like the residue mass, size, charge, hydrophobicity, C-beta density, and residue flexibility [**?**, 1]. Such methods may also rely on known protein structures in order to incorporate properties such as local secondary structure and solvent accessibility. In the absence of structural information, however, computational prediction accuracy is low, and full three-dimensional structures have been experimentally solved for a tiny fraction of proteins [2, 3] (<35% of human protein coding genes have structures deposited in PDB [4]). With few exceptions, therefore, these methods have not been applied on a genome-wide scale.

Despite the common framing that "sequence determines structure which determines function", single nucleotide polymorphisms (SNPs) can alter function while leaving the protein structure essentially unchanged. For example, intrinsically disordered proteins (IDPs) lack unique structure yet contain SNPs in their disordered regions that are associated with signaling [5, 6] and aggregation disorders [7, 8], including psychiatric [**?**], neurodegenerative [9], and aging-related decline [10], and IDPs have key functions in many critical biological pathways [11–15]. IDPs tend to evolve rapidly, with poor conservation across species [16], and harbor many common polymorphisms in global populations today. For a large class of common human polymorphisms, therefore, it is

particularly difficult to predict whether a variant has an effect, as "structural" impacts are not applicable and homologs in model organisms may be unreliable if they exist at all.

Although not all functional proteins have well-defined structures, most (including IDPs) are intrinsically modular. A given protein may include multiple secondary structure elements, ordered and disordered regions, transmembrane and soluble portions, or stretches of highly charged regions followed by stretches of highly hydrophobic regions. Incorporating this intrinsic modularity into bioinformatics "variant-to-function" prediction methods is an ongoing endeavor. Most bioinformatics prediction methods that use sequence context define the local sequence using a window of constant length, centered around the mutation. This definition of sequence context can weaken the predictive ability, particularly for SNPs near the boundary between adjacent protein modules with distinct functional roles. For example, if the mutation is at the terminus of an alpha helix, near the interface between two domains, or in a domain that is shorter than the window length, residues that are outside the module containing the SNP will still be included in the analysis and necessarily weaken predictive ability.

In a previous study [17], we developed a sequence-based approach for identifying modularity in order to analyze molecular dynamics simulation data of the long disordered prodomain of brain derived neurotrophic factor (BDNF). We identified contiguous stretches of residues with similar hydrophobicity, also known as "blobs". Blobs may contain secondary structure elements, but are not required to do so. The dynamics of residues in a hydrophobic blob might not be correlated by a rigid backbone structure, but will still be correlated by the cooperativity induced by the hydrophobic effect. This approach allowed us to identify the conformational signal of the disease-associated Val66Met SNP within a large and very noisy data set, and thus determine a mechanism for the mutation's effects.

In the present study, we apply this approach to a database of protein sequences in order to determine the extent to which hydrophobic-based sequence segmentation captures functional modules. The overall approach involves three steps: 1) segmentation of the protein into the blobs based on contiguous hydrophobicity that will define local sequences ("blobulation"), 2) characterization of the blobs based on various physiochemical properties, including hydrophobicity, and 3) testing whether SNPs with "known" functional impact are enriched or depleted genome-wide in blobs with well-defined characterizations. The two blob characterizations we consider are the mean hydrophobicity class and the Das-Pappu globular class based on the fractions of positively and negatively charged residues.

While we are not aware of a similar approach applied to generic proteins, hydrophobic blobs are analogous to the aggregation "hot spots" identified by tools such as AGGRESCAN [18], ProA [19], and Zyggregator [20]. Detection of such hot spots from sequence is of considerable interest [21–24], because protein aggregation is implicated in several pathological conditions such as Alzheimer's, Parkinson's, prion diseases and diabetes. [21, 25, 26]. Among various sequence properties identified as aggregation characteristics (hydrophobicity, charge, alternating patterns of hydrophobic and hydrophilic stretches, secondary structure propensity, packing density), residue hydrophobicity has been fairly common. Here we also specifically test whether the blobulation approach detects enrichment of disease causing SNPs in the hydrophobic blobs of aggregation-prone proteins.

Results

Blobs containing disease-associating SNPs are more likely to be hydrophobic and/or weak polyampholytes

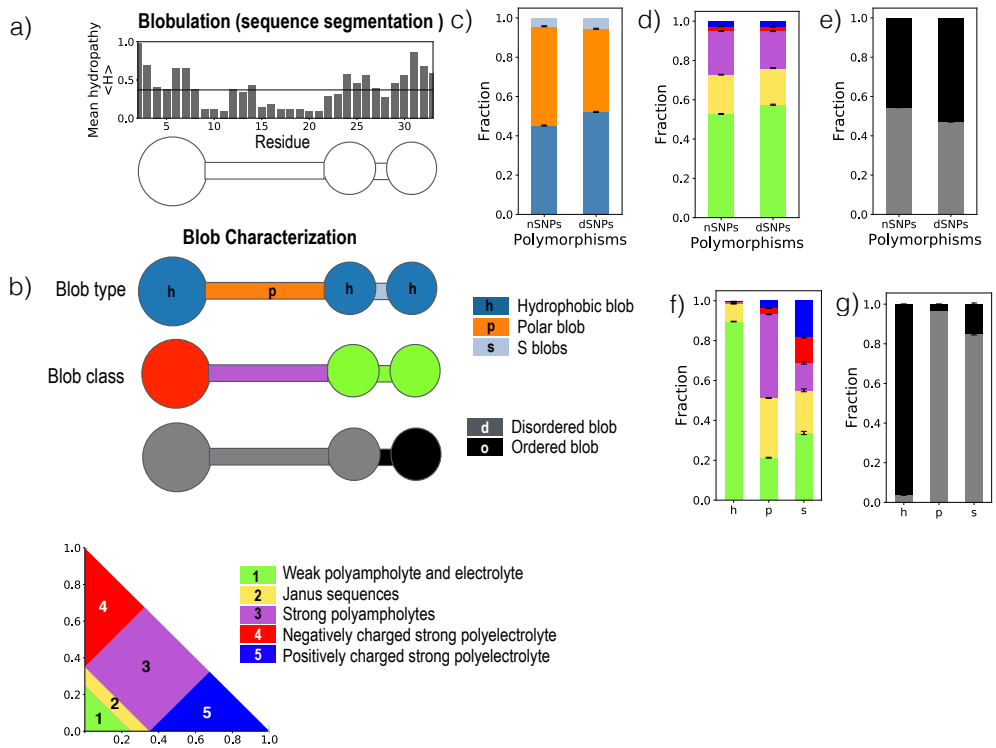

Fig 1. Blobulation methodology and distribution of blob characteristics in dSNPs and nSNPs . a) Cartoon representation of the blobulation approach. First, hydrophobicity (average of 3 residue moving window) is obtained for each residue. Blobs are identified as contiguous stretches of at least 4 residues (unless otherwise noted) in which each residue is above, or each residue is below the hydrophobicity cutoff (black line). check please, Ruchi b) Cartoon representation in which the shape of the blobs is determined by hydrophobicity class: h blobs are shown as circles and p "blobs" are shown as rectangles. Blobs are colored according to hydrophobicity class(top) and charge class(bottom) c) Fraction of nSNPs or dSNPs that are found in either h blobs (orange) or p blobs (blue). d) Fraction of nSNPs or dSNPs that are found in blobs of each charge class. (e). Fraction of h or p blobs that are found in each charge class. The distribution in c, d and e are plotted with a cutoff and minimum blob length of 0.4 and 4 respectively. The effect of these parameters on proportion of SNPs in blob type is further discussed in Fig 3b. If enrichment or depletion in dSNPs is significant ($p < 5 \times 10^{-3}$) it is annotated with a star. Errors bars in (c)-(e) represent one standard error for multinomial distributed data.1. Das Pappu should be its own panel (probably panel b) and should have its own line in the caption. 2. Update blob type and blob class labels. 3. Blue/orange Coloring is still backwards between Panel C and B. 4. Plot in panel e should be before c and d.

In order to test blobulation as a meaningful approach for analyzing sequences, we tested whether any blob characterization was correlated with a higher incidence of disease-causing mutations. For each SNP, we applied the blobulation approach on the

entire protein sequence (Fig 1a) as described in Methods, and then further analyzed the blob containing the SNP. Specifically, we measured the enrichment of disease-associated SNPs (dSNPs) relative to non-disease-associated SNPs (nSNPs) for blobs characterized by hydrophobicity class and charge class (SNP dataset obtained from Uniprot; see Methods). Unless otherwise noted, dSNPs are tested for enrichment relative to the expectation set by nSNPs . For example, the phrase "dSNPs are enriched in X blobs" means that dSNPs are found at a higher rate in blobs of type X than are nSNPs . Fig 1c compares the distribution of blob hydrophobicity classamong dSNPs and nSNPs . We find that dSNPs are enriched in h blobs by about 1.15 fold: the fraction of dSNPs in h blobs is (52%) while the fraction of nSNPs in h blobs is about 45%. This is consistent with our previous results from simulations of the long disordered BDNF prodomain, in which h blobs interacted frequently with each other, while p blobs did not form any specific interactions with the rest of the sequence.

In addition to classification by hydrophobicity class, the blobs were also classified according to charge class (Fig 1b), which is the predicted globular phase based on the fraction of positive and negative charges (the Das-Pappu [**?**] phase). Possible values of the blob charge class are 1 (Weak polyampholyte), 2 (Janus or boundary region), 3 (Strong polyampholyte), 4 (Negatively-charged strong polyelectrolyte), and 5 (Positively-charged strong polyelectrolyte). Sequences in class 1 have a low fraction of positively charged residues, and nearly all structured proteins fall in class 1. In contrast to structured proteins, IDPs can be found in all five Das-Pappu phases, particularly classes 3, 4, and 5. Therefore the Das-Pappu charge classcharacterization provides a natural metric for distinguishing among sub-categories of IDPs with different functional behavior.

The blob hydrophobicity class and charge class are fundamentally correlated; while blob charge class does not explicitly consider hydrophobicity, increasing the number of charged residues will reduce the average hydrophobicity of a blob. The extent of this correlation is shown in (Fig 1e), which breaks down the fraction of h and p blobs that fall in each Das-Pappu charge class. As expected, most h blobs (89%) fall in class 1 (weak polyampholyte), followed by 8% in class 2 (Janus). The p blobs are more evenly distributed across classes, with the highest fraction (40%) classified as strong polyelectrolytes.

The distribution of dSNPs and nSNPs across blob charge class are shown in Fig 1d. More than half of both dSNPs and nSNPs are found in blobs that are weak polyampholytes. dSNPs are slightly enriched (1.09 fold) for weak polyampholyte blobs. Due to the high correlation for h blobs and the additional sequence information encapsulated for p blobs, we expected that blob charge class would be at least as sensitive, if not more sensitive, than blob hydrophobicity class for identifying functional protein segments, and thus have generally higher enrichment for disease-associated SNPs. Surprisingly, we found that the maximum charge class enrichment (1.09) is slightly less than the enrichment found for h blobs (1.15 fold). Although most hydrophobic blobs are weak polyampholytes, this difference in enrichment suggests that hydrophobicity (rather than simple lack of charge) may be more indicative of sensitivity to mutation. Strong polyampholytes are depleted among dSNPs (0.87 fold), which may indicate a relative decrease in blob-blob interactions, since strong polyampholytes tend to be highly solvent exposed.

Disease-associated SNPs are enriched for mutations that change blob classifications

Since blob properties like hydrophobicity class and charge class are dependent upon blob sequence, a SNP may directly result in a change of blob classification. Such

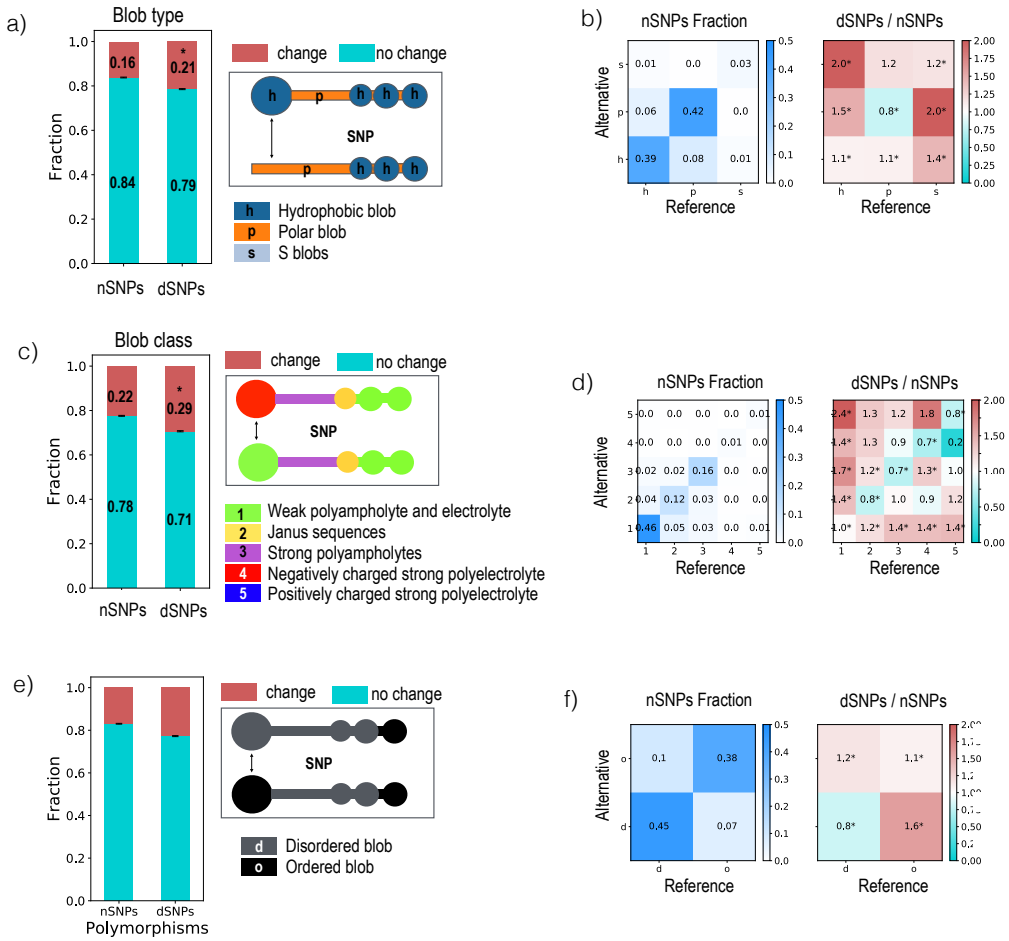

Fig 2. Frequency of SNPs that change blob class. a) Fraction of nSNPs or dSNPs that either do (pink) or do not (blue) induce a change in hydrophobicity class. b) Fractions (as in a) broken down further by the blob hydrophobicity class before (x axis) and after (y axis) the mutation. Grid boxes are colored according to the overall fraction of corresponding nSNPs (left) and the corresponding enrichment ratios (dSNPs /nSNPs), with the specific values shown in the boxes. c) As in a, but for transitions in charge class. d) As in b, but for transitions in charge class. Fewer than ? % of SNPs involve class 4 and 5, and these data are not shown for readability. For all panels, significant enrichment or depletion in dSNPs ($p < 5 \times 10^{-3}$) is annotated with a star. Errors bars in (a) and (c) represent one standard error for multinomial distributed data.

transitions will be far more common in shorter blobs, where each residue contributes more to the overall blob assignment. We observe that about 17% of the nSNPs and 22% of dSNPs involve a blob-hydrophobicity class change, yielding a 1.3 fold enrichment for blob hydrophobicity class changes (Fig 2a).

We calculated the frequency of each of the four possible blob-hydrophobicity class transitions: h $\rightarrow$ p, h $\rightarrow$ h, p $\rightarrow$ h, and p $\rightarrow$ p (Fig 2b). We find that 7% of SNPs in the dataset cause h $\rightarrow$ p blob transitions, and these yield the maximum enrichment (1.6 fold) among dSNPs . These results suggest a particularly high likelihood of functional impact upon introduction of a less hydrophobic residue in a generic hydrophobic sequence (for example, introduction of a charged residue into a buried position).

Slightly more SNPs induce the reverse transition (p $\rightarrow$ h), but dSNPs have only 1.1 fold enrichment for this transition. Additionally, SNPs in p blobs causing no blob hydrophobicity class change are depleted among disease-associated SNPs. Overall these results are consistent with the increased mutational sensitivity of hydrophobic (and typically buried) blobs that is shown in Figure 1.

Mutations that reverse residue charge would be expected to have particularly strong functional effects, but would not affect the blob hydrophobicity class. They could, however, affect the blob charge class. Mutations involving a neutral residue and a charged residue could also directly affect the blob charge class, either directly, by changing the fraction of positive/negative residues in the original blob or indirectly, by affecting the blob segmentation. No indirect effect for hydrophobicity class? needs discussion

The frequency of SNP-induced changes in blob charge class class is shown in Fig 2d, for blobs in category 1 (weak polyampholyte), category 2 (Janus), or category 3 (strong polyampholyte). Transitions involving categories 4 or 5 (positively or negatively-charged strong polyelectrolyte) represented fewer than how much? % of the total transitions. Disease-associated SNPs are enriched for all mutations that change blob charge class and either unenriched or weakly depleted for mutations that do not change blob charge class. The degree of enrichment is largest for transitions between weak and strong polyampholytes, so $1 \rightarrow 3 > 1 \rightarrow 2$ and $3 \rightarrow 1 > 3 \rightarrow 2$. Disease-associated SNPs are most strongly enriched (1.8 fold) for SNPS that switch a weak polyampholyte (1) blob to a strong polyampholyte (3) blob were the most strongly enriched for disease-association. This is consistent with evidence that the severity of a SNP on protein function is somewhat weakly? correlated with the physicochemical difference between the original amino acid and the missense variant [27].Can we make this statement stronger, use more recent reference?

These results suggest that transitions in blob charge class may be particularly useful for predicting whether variants have a functional effect. While generic transitions of blob hydrophobicity class or charge class yield an equivalent amount of enrichment (1.3 fold) for disease-association, more SNPs in this dataset induce a blob charge class transition than a hydrophobicity class transition (29% vs 21%, respectively). Furthermore, charge class distinguishes between lower and higher enrichment transitions, since most $1\rightarrow 2$ mutations (1.4 fold enrichment) and $1\rightarrow 3$ mutations (1.8 fold enrichment) will also count as h$\rightarrow$ p mutations (1.6 fold enrichment). This result suggests that for SNPs that change the blob hydrophobicity class, a further split into those with weaker or stronger effects on charge class provides greater resolution of functional effects.

Blobulation yields higher enrichment values and more meaningful trends than fixed-length moving windows

The blobulation approach is a particular systematic approach for identifying segments/blobs in the protein sequence. Two parameters (minimum blob length and hydrophobicity cutoff) are required to define blob edges. The degree of enrichment for disease-association is expected to be dependent upon both of these parameters, and we ran the same enrichment tests as in Figure 1 for a wide range of minimum blob lengths and hydrophobicity cutoff values (Fig 3a). The highest enrichment is found for blobs that have a contiguous stretch of at least 20 residues with hydrophobicity greater than 0.4; check valuesmore than twice as many dSNPs than nSNPs are located in such blobs. We should provide these as tables in SI heatmap values?. This result indicates that hydrophobicity class-based sequence segmentation could be particularly useful for assessing the riskiness of SNPs located in long and very hydrophobic sequences.

In order to compare the effectiveness of blobulation to a moving window approach,

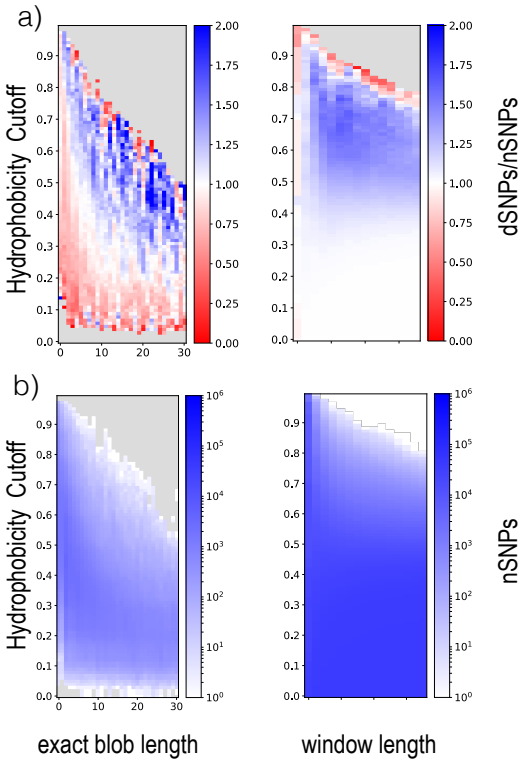

Fig 3. Effect of segmentation approach, length, and hydrophobicity cutoff on calculated enrichment of dSNPs in hydrophobic segments. a) Enrichment of dSNPs for hydrophobic segments using a blobulation approach (left) and moving window approach (right). Blobulation defines a hydrophobic segment (or h blob) as a contiguous stretch of a minimum length in which every residue has hydrophobicity greater than the cutoff. In contrast, the moving window approach defines a hydrophobic segment as a stretch of residues of a specific length with a mean hydrophobicity that is greater than the cutoff. 1) Let's discuss non-monotonicity. Is it real? 2) x axis should be minimum blob length 3) title label is confusing, let's discuss. Each bin is colored according to the fraction of dSNPs in hydrophobic segments divided by the analogous nSNP fraction, with the scale shown in the color bar. No data is shown for parameters in which there are fewer than ten SNPs assigned to hydrophobic segments (gray regions). b) Effect of segmentation parameters on the overall populations of nSNPs in hydrophobic segments, using the blobulation approach (left) and moving window approach (right). Each bin is colored according to the fraction of nSNPs that fall in hydrophobic segments.should show log(fraction).Matt says: i suggest that we make both the x and y bins larger. may alleviate some of the small numbers issues and clean up some of the visible noise. agree that in b) should show log(fraction). And the panels need some form of title/label to show which one is blobulation and which is sliding windows. Also, need to state the bin sizes..

we ran analogous enrichment tests for sliding windows across all protein sequences. Two analogous parameters (window length and mean hydrophobicity cutoff) are used for defining hydrophobic regions in the moving window approach, with the resulting enrichment for a range of these values also shown in Fig 3a. Compared to the blobulation approach, the enrichment detected using the moving window approach is much less sensitive to segment length for windows beyond 6 residues (consider the tilted

color bands for the blobulation approach vs the horizontal color bands for moving windows in Figure 3a).

While there is no "standard" window size, most SNP prediction programs use a window size in the range of 1-21 residues. The window size is chosen to balance concerns that small window sizes may not accurately capture the "local" sequence [28–30] whereas long window sizes can increase signal to noise ratio [31]. The enrichment insensitivity to window size was thus surprising, but the expected noise was likely lost due to averaging across the proteome. As shown in Fig 3b, the overall fraction of hydrophobic segments for mild or moderate hydrophobicity cutoffs was also insensitive to window size, confirming that the dataset we use here is sufficiently large to average out large window noise. For a single SNP, considering too many residues far from the SNP will still contribute significant noise to the prediction.

Overall, use of the moving window approach yields two-fold enrichment for only a few, scattered parameter values with very few qualifying sequences. In contrast, blobulation yielded greater than two-fold enrichment for a well-defined set of parameters (hydrophobicity cutoff > what? and minimum blob size > what?. This observation supports our hypothesis that blobulation provides a more meaningful and less noisy approach to protein segmentation than use of a fixed-length moving window.

Disease-causing SNPs in aggregating proteins are particularly enriched in hydrophobic blobs

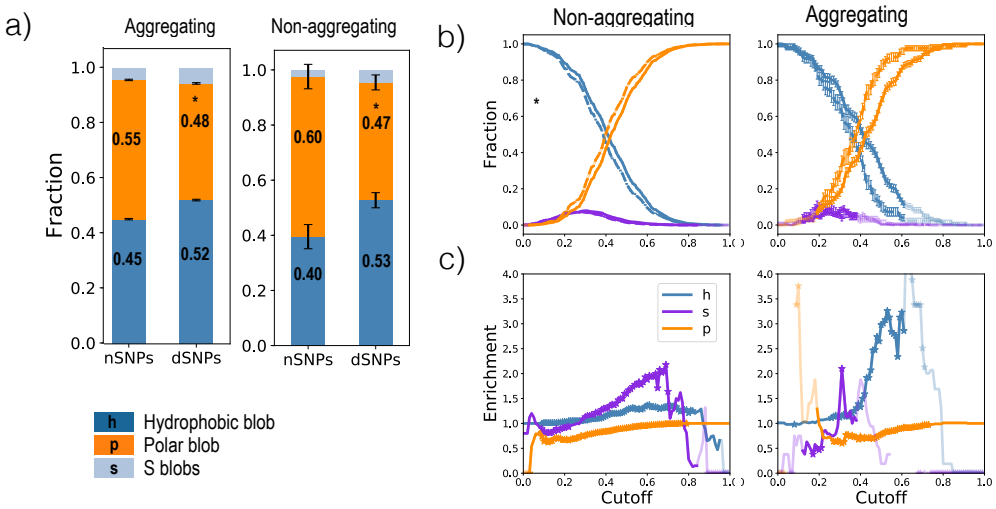

Fig 4. Blob hydrophobicity for SNPs in aggregating proteins. a) Fraction of SNPs that fall in h or p blobs in known-aggregating proteins (left) and all others in the dataset (right). b) The population of nSNPs (dashed) and dSNPs (solid) in h (blue) or p (orange) blobs for a given hydrophobicity cutoff, in known-aggregating proteins (left) and all others (right). dSNPs are significantly enriched in hydrophobic blobs for both aggregating and non-aggregating proteins ($p < 5 \times 10^{-3}$, indicated by a star). Error bars represent one standard error for multinomial distributed data. c) The corresponding enrichment (dSNP fractions divided by nSNP fractions) for a range of hydrophobicity cutoffs. All blob assignments used a minimum blob length of 4, so the values for non-aggregating proteins are equivalent to the $x = 4$ column in the left heat map of Figure 3A.

In order to test the hypothesis that the blob hydrophobicity classis a stronger

predictor of the functional impact of a SNP in aggregating proteins when compared to non-aggregating proteins, we separated known aggregating proteins out of the dataset. Proteins involved in formation of extracellular amyloid deposits or intracellular inclusions with amyloid-like characteristics are labelled as "Aggregating proteins" (28 proteins, 124 nSNPs, 330 dSNPs) and all the remaining proteins are labelled as "Non-aggregating proteins".

As shown in Fig 4a, with the mild hydrophobicity cutoff of 0.4 used in Figs(1 and 2), dSNPs in aggregating proteins have a 1.33 fold enrichment in h blobs. In comparison, dSNPs in non-aggregating proteins show a 1.15 fold enrichment in h blobs. This is consistent with the well-established hydrophobicity of aggregating proteins, including the hydrophobicity of amyloidogenic sequences [32, 33], and further supports the hydrophobic regions of those proteins as particularly sensitive to mutation. Need a sentence comparing to hot spot detection, particularly whether residues with high hydrophobicity also have high aggregation tendencies. Sentences in the blue paragraph just below are too vague for this section, but circle around what we need.

Among various sequence properties identified as aggregation characteristics (hydrophobicity, charge, alternating patterns of hydrophobic and hydrophilic stretches, secondary structure propensity, packing density), residue hydrophobicity has been fairly common. Aggregation "hot spots" are often found in the hydrophobic core of the proteins. Since, hydrophobicity is one of the few residue properties used for identifying aggregation "hot spots", it will likely be a stronger determinant for the functional impact of SNPs in theses proteins when compared with the functional impact of SNPs in non-aggregating proteins. Finding SNPs in aggregation "hot spots" is similar to finding SNPs in the functional/active sites of proteins, since these SNPs can directly affect the proteins' aggregation propensity and function. comment repeated motivation

More significantly, the enrichment for dSNPs in the h blobs of aggregating proteins increases steadily as the requirements for h blobs are made stricter, reaching a maximum of 3 fold exact number? enrichment for blobs meeting a 0.6 hydrophobicity cutoff (Fig 4b). This is consistent with previous observations that the in vitro aggregation rate of unstructured polypeptide chains is proportional to hydrophobicity and inversely proportional to net charge [34]. In contrast, non-aggregating proteins are relatively insensitive to hydrophobic cutoff, reaching a maximum enrichment of less than 1.5 fold.

Discussion

In the present work, we have tested whether the protein "blobulation" approach we developed in [17] for a specific disordered protein is a meaningful approach for identifying segments across generic proteins. We found that

1. Overall, disease-associated mutations are weakly enriched (1.15 fold) in hydrophobic blobs.

2. Disease-associated mutations are moderately enriched for mutations that cause transitions in blob hydrophobicity class (up to 1.6 fold) and strongly enriched for mutations that cause certain transitions in blob charge class (1.8 fold).

3. Enrichment of disease-associated mutations in hydrophobic blobs increases with the strictness of the hydrophobic blob criteria.

4. SNPs in variable-length hydrophobic blobs were more strongly associated with disease than SNPs in fixed-length hydrophobic moving windows, regardless of the parameters used.

5. Disease-causing SNPs in aggregating proteins were particularly enriched in hydrophobic blobs. Disease-associated SNPS were more than 3 fold enriched in h blobs that met the strictest criteria.

These results support the use of blobulation as a simple method for segmenting proteins, requiring only the protein amino acid sequence and two parameters (minimum blob length and hydrophobicity cutoff). Once blobs are identified, they can be characterized on any property of interest, including mean hydrophobicity, net charge per residue, fraction of polar residues, fraction of disordered residues, etc. The present results also support a role for hydrophobicity in determining the sensitivity of a given protein region to disease-causing mutations.

We anticipate that these results may further efforts to assess functional significance of SNPs for which disease-causality is uncertain. Despite the wealth of genetics information presently available, the complex nature of many phenotypes makes the identification of causal variants difficult [35]. Widely used methods to find genotype/phenotype associations, such as pedigree-based linkage studies or population-based association studies, have identified thousands of associations with a wide spectrum of phenotypes and diseases [36–38]. Association of a SNP with a disease does not necessarily indicate causality, [39–42] motivating the development multiple complementary approaches for predicting the functional effects of SNPs.

All protein level methods rely on some form of residue characterization for such predictions. In addition to physicochemical properties [1, 27, 43–45] similar to the hydrophobicity class and charge class considered in the present work, these may include evolutionary conservation [1, 27, 43, 46, 46–49] and structural propensities [45, 50–55]. Methods relying on physiochemical properties are uniquely suited for application to disordered sequences and weakly conserved sequences, and may be compared directly against in vitro experimental results for mechanistic consistency. Yet many are still limited by a reliance on structural information for predictive accuracy.

While structural data implicitly provides insight into the segmentation of a protein sequence, we show here that segmentation may be meaningfully estimated from the sequence alone. Based on the present results, we suggest that structureless methods may recapture some accuracy of structure-based methods by introducing sequence-based segmentation. While we propose blobulation as a simple approach for doing so, in principle, secondary structure prediction could be used instead. For many proteins, however, this approach would be frequently unfeasible and overly restrictive. In addition to the challenges inherent in predicting secondary structure (which may be highly sensitive to the local protein environment or simply non-existent), many secondary structure prediction methods require alignment to a homologous sequence with known structure. [56–60]. This is essential for secondary structure predictors to achieve their primary goal: distinguishing *between* secondary structures. Segmentation, however, only requires knowing where the segment begins and ends.

SNP-prediction methods often test residue properties for relevance by machine learning algorithms such as neural networks [1, 44, 61], Hidden Markov Models [46, 62], Random forests [43, 55] or Support vector Machines [49, 63–65]). The machine learning methods derive their decision rules based on training datasets of annotated mutations, and therefore are sensitive to the training set used. Machine learning methods also rely on a large number of features, which can obscure the biological relevance of any given feature. Here we have instead performed hypothesis-driven tests for enrichment of certain features in putatively causal SNPs, based on the principle that those features that are most relevant for evaluating SNPs should be detectable by simple enrichment tests. Nonetheless, the characteristics of the SNP-containing blob should be straightforward to add as an input to a machine-learning model. In particular, it could be a tractable approach for combining blob-level information with protein-level

information, as we did for the particular case of aggregating proteins.

1 Methods

Ruchi, please give attention to this methods section, it is very incomplete.

Datasets

The list of all missense SNPs annotated in human UniProtKB/Swiss-Prot entries was obtained from http://beta.uniprot.org/docs/humsavar (last Release: 17-Jun-2020) [66]. This manually curated catalog contains missense mutations on the most common isoform of the given protein and does not contain frameshift and nonsense mutations. A SNP is annotated as 'Disease-associated (dSNPs)' or 'Non Disease-associated (nSNPs)' depending on if it is implicated in disease or not according to literature reports. nSNPs is also used to describe rare SNPs as well as polymorphisms that have an effect on protein function, but with no resulting clinical phenotype (functional polymorphisms) [66]. A total of 69,675 SNPs were analyzed from 12,507 proteins. Among the total missense SNPs found in the database, 30,227 (43.4%) are dSNPs while the remaining 39,448 (56.6%) are nSNPs .

A list of 28 proteins (P02647, P06727, *P02655, P05067, Q99700, P61769, P01258, *P17927, P07320, P01034, *P35637, P06396, Q9NX55, P10997, P08069, *P01308, P02788, *P61626, P10636, Q08431, P01160, P04156, P11686, P37840, P00441, *Q13148, *Q15582, P02766) was labelled as "Aggregating proteins". These proteins are involved in formation of extracellular amyloid deposits or intracellular inclusions with amyloid-like characteristics [67].

Blobulation

Mean hydrophobicity ($\langle H \rangle$) at each residue is defined as the average Kyte-Dolittle [68] score with a window size of three residues, scaled to fit between 0 and 1. For each protein, four (unless otherwise noted) or more contiguous residues above the cutoff (0.4) were identified as forming a h blob. Among the remaining residues contiguous stretch of four (unless otherwise noted) or more residues is classified as p blob otherwise as s residues.

Statistical analysis

Binomial test was used for calculating the fold enrichment. Any enrichment or depletion in dSNPs is significant if $p < 5 \times 10^{-3}$.

Acknowledgments

References

1. Hecht M, Bromberg Y, Rost B. Better prediction of functional effects for sequence variants. BMC Genomics. 2015;16(S8). doi:10.1186/1471-2164-16-s8-s1.

2. Mir S, Alhroub Y, Anyango S, Armstrong DR, Berrisford JM, Clark AR, et al. PDBe: towards reusable data delivery infrastructure at protein data bank in Europe. Nucleic Acids Research. 2017;46(D1):D486–D492. doi:10.1093/nar/gkx1070.

3. Peter W Rose AACBARBCHCLDCJMDSDZFRKGDSGBHTKRLEPCRASRC-SYPTYVMVJDWJWHYJYYCZHMBSKB Andreas Prlić. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Research. 2016;doi:10.1093/analys/anz030.

4. Prlić A, Kalro T, Bhattacharya R, Christie C, Burley SK, Rose PW. Integrating genomic information with protein sequence and 3D atomic level structure at the RCSB protein data bank. Bioinformatics. 2016;32(24):3833–3835. doi:10.1093/bioinformatics/btw547.

5. Uversky VN, Iakoucheva LM, Dunker AK. Protein Disorder and Human Genetic Disease;.

6. Deiana A, Giansanti A. Variants of intrinsic disorder in the human proteome;.

7. Patel A, Lee HO, Jawerth L, Maharana S, Jahnel M, Hein MY, et al. A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. Cell. 2015;162(5):1066–1077. doi:10.1016/j.cell.2015.07.047.

8. Uversky VN. Intrinsically disordered proteins and their (disordered) proteomes in neurodegenerative disorders. Frontiers in Aging Neuroscience. 2015;7. doi:10.3389/fnagi.2015.00018.

9. Weinreb PH, Zhen W, Poon AW, Conway KA, Lansbury PT. NACP, A Protein Implicated in Alzheimer's Disease and Learning, Is Natively Unfolded†. Biochemistry. 1996;35(43):13709–13715. doi:10.1021/bi961799n.

10. Cuanalo-Contreras K, Mukherjee A, Soto C. Role of Protein Misfolding and Proteostasis Deficiency in Protein Misfolding Diseases and Aging. International Journal of Cell Biology. 2013;2013:1–10. doi:10.1155/2013/638083.

11. Uversky VN. Unusual biophysics of intrinsically disordered proteins. Biochim Biophys Acta. 2013;1834(5):932–51. doi:10.1016/j.bbapap.2012.12.008.

12. Panchenko AR, Babu MM. Editorial overview: Linking protein sequence and structural changes to function in the era of next-generation sequencing. Curr Opin Struct Biol. 2015;32:viii–x. doi:10.1016/j.sbi.2015.06.005.

13. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. J Mol Biol. 2004;337(3):635–645. doi:10.1016/j.jmb.2004.02.002.

14. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol. 2005;6(3):197–208. doi:10.1038/nrm1589.

15. Uversky VN. Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. Front Phys. 2019;7:10. doi:10.3389/fphy.2019.00010.

16. Brown CJ, Johnson AK, Dunker AK, Daughdrill GW. Evolution and disorder. Current opinion in structural biology. 2011;21(3):441–446.

17. Lohia R, Salari R, Brannigan G. Sequence specificity despite intrinsic disorder: how a disease-associated Val/Met polymorphism rearranges tertiary interactions in a long disordered protein;doi:10.26434/chemrxiv.8135777.

18. Conchillo-Solé O, de Groot NS, Avilés FX, Vendrell J, Daura X, Ventura S. AGGRESCAN: a server for the prediction and evaluation of" hot spots" of aggregation in polypeptides. BMC bioinformatics. 2007;8(1):65. doi:10.1186/1471-2105-8-65.

19. Fang Y, Gao S, Tai D, Middaugh CR, Fang J. Identification of properties important to protein aggregation using feature selection. 2013;14. doi:10.1186/1471-2105-14-314.

20. Tartaglia GG, Vendruscolo M. The Zyggregator method for predicting protein aggregation propensities. 2008;37:1395. doi:10.1039/b706784b.

21. Pallarés I, Ventura S. Advances in the Prediction of Protein Aggregation Propensity. Current Medicinal Chemistry. 2019;26(21):3911–3920. doi:10.2174/0929867324666170705121754.

22. Maurer-Stroh S, Debulpaep M, Kuemmerer N, de la Paz ML, Martins IC, Reumers J, et al. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nature Methods. 2010;7(3):237–242. doi:10.1038/nmeth.1432.

23. Tartaglia GG, Cavalli A, Pellarin R, Caflisch A. Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. Protein Science. 2005;14(10):2723–2734. doi:10.1110/ps.051471205.

24. Pawar AP, DuBay KF, Zurdo J, Chiti F, Vendruscolo M, Dobson CM. Prediction of "Aggregation-prone" and "Aggregation-susceptible" Regions in Proteins Associated with Neurodegenerative Diseases. Journal of Molecular Biology. 2005;350(2):379–392. doi:10.1016/j.jmb.2005.04.016.

25. Dobson CM. Protein-misfolding diseases: Getting out of shape. Nature. 2002;418(6899):729–730.

26. Cohen F. Kelly, JW". Therapeutic approaches to protein-misfolding diseases Nature. 2003;426:905–909.

27. Stone EA. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Research. 2005;15(7):978–986. doi:10.1101/gr.3804205.

28. Chen K, Kurgan L, Ruan J. Optimization of the Sliding Window Size for Protein Structure Prediction. In: 2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology. IEEE; 2006.

29. Schlessinger A, Rost B. Protein flexibility and rigidity predicted from sequence. Proteins: Structure, Function, and Bioinformatics. 2005;61(1):115–126. doi:10.1002/prot.20587.

30. Sander O, Sommer I, Lengauer T. Local protein structure prediction using discriminative models. BMC Bioinformatics. 2006;7(1):14. doi:10.1186/1471-2105-7-14.

31. Park Y, Hayat S, Helms V. Prediction of the burial status of transmembrane residues of helical membrane proteins. BMC Bioinformatics. 2007;8(1):302. doi:10.1186/1471-2105-8-302.

32. Tzotzos S, Doig AJ. Amyloidogenic sequences in native protein structures. Protein Science. 2010;19(2):327–348. doi:10.1002/pro.314.

33. Abskharon R, Wang F, Wohlkonig A, Ruan J, Soror S, Giachin G, et al. Structural evidence for the critical role of the prion protein hydrophobic region in forming an infectious prion. PLOS Pathogens. 2019;15(12):e1008139. doi:10.1371/journal.ppat.1008139.

34. DuBay KF, Pawar AP, Chiti F, Zurdo J, Dobson CM, Vendruscolo M. Prediction of the Absolute Aggregation Rates of Amyloidogenic Polypeptide Chains. Journal of Molecular Biology. 2004;341(5):1317–1326. doi:10.1016/j.jmb.2004.06.043.

35. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, et al. A brief history of human disease genetics. Nature. 2020;577(7789):179–189. doi:10.1038/s41586-019-1879-7.

36. Buniello A, MacArthur JA, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. 2019;47:D1005–D1012. doi:10.1093/nar/gky1120.

37. Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. 2018;102:717–730. doi:10.1016/j.ajhg.2018.04.002.

38. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. 2019;20:467–484. doi:10.1038/s41576-019-0127-1.

39. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. 2017;169:1177–1186. doi:10.1016/j.cell.2017.05.038.

40. Goldstein DB. Common Genetic Variation and Human Traits. 2009;360:1696–1698. doi:10.1056/nejmp0806284.

41. McClellan J, King MC. Genetic Heterogeneity in Human Disease. 2010;141:210–217. doi:10.1016/j.cell.2010.03.032.

42. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508(7497):469–476. doi:10.1038/nature13127.

43. Niroula A, Urolagin S, Vihinen M. PON-P2: Prediction Method for Fast and Reliable Identification of Harmful Variants. PLOS ONE. 2015;10(2):e0117380. doi:10.1371/journal.pone.0117380.

44. López-Ferrando V, Gazzo A, de la Cruz X, Orozco M, Gelpí JL. PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. Nucleic Acids Research. 2017;45(W1):W222–W228. doi:10.1093/nar/gkx313.

45. Popov P, Bizin I, Gromiha M, A K, Frishman D. Prediction of disease-associated mutations in the transmembrane regions of proteins with known 3D structure. PLOS ONE. 2019;14(7):e0219452. doi:10.1371/journal.pone.0219452.

46. Thomas PD, Kejariwal A. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: Evolutionary evidence for differences in molecular effects. Proceedings of the National Academy of Sciences. 2004;101(43):15398–15403. doi:10.1073/pnas.0404380101.

47. Ng PC, Henikoff S. Predicting Deleterious Amino Acid Substitutions. Genome Research. 2001;11(5):863–874. doi:10.1101/gr.176601.

48. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS ONE. 2012;7(10):e46688. doi:10.1371/journal.pone.0046688.

49. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. 2006;22(22):2729–2734. doi:10.1093/bioinformatics/btl423.

50. Iqbal S, Jespersen JB, Perez-Palma E, May P, Hoksza D, Heyne HO, et al. Insights into protein structural, physicochemical, and functional consequences of missense variants in 1,330 disease-associated human genes. bioRxiv. 2019; p. 693259. doi:10.1101/693259.

51. Ittisoponpisan S, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE. Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? Journal of Molecular Biology. 2019;431(11):2197–2212. doi:10.1016/j.jmb.2019.04.009.

52. Capriotti E, Fariselli P, Casadio R. A neural-network-based method for predicting protein stability changes upon single point mutations. Bioinformatics. 2004;20(Suppl 1):i63–i68. doi:10.1093/bioinformatics/bth928.

53. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Research. 2005;33(Web Server):W306–W310. doi:10.1093/nar/gki375.

54. Parthiban V, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Research. 2006;34(Web Server):W239–W242. doi:10.1093/nar/gkl190.

55. Wainreb G, Ashkenazy H, Bromberg Y, Starovolsky-Shitrit A, Haliloglu T, Ruppin E, et al. MuD: an interactive web server for the prediction of non-neutral substitutions using protein structural data. Nucleic Acids Research. 2010;38(suppl_2):W523–W528. doi:10.1093/nar/gkq528.

56. Yang Y, Gao J, Wang J, Heffernan R, Hanson J, Paliwal K, et al. Sixty-five years of the long march in protein secondary structure prediction: the final stretch? 2018; p. bbw129. doi:10.1093/bib/bbw129.

57. Zhang B, Li J, Lü Q. Prediction of 8-state protein secondary structures by a novel deep learning architecture. 2018;19. doi:10.1186/s12859-018-2280-5.

58. Wang Y, Mao H, Yi Z. Protein secondary structure prediction by using deep learning method. 2017;118:115–123. doi:10.1016/j.knosys.2016.11.015.

59. Aydin Z, Altunbasak Y, Borodovsky M. Protein secondary structure prediction with semi Markov HMMs;.

60. Ma Y, Liu Y, Cheng J. Protein Secondary Structure Prediction Based on Data Partition and Semi-Random Subspace Method. 2018;8. doi:10.1038/s41598-018-28084-8.

61. Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Research. 2007;35(11):3823–3835. doi:10.1093/nar/gkm238.

62. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. Human Mutation. 2012;34(1):57–65. doi:10.1002/humu.22225.

63. Yue P, Li Z, Moult J. Loss of Protein Structure Stability as a Major Causative Factor in Monogenic Disease. Journal of Molecular Biology. 2005;353(2):459–473. doi:10.1016/j.jmb.2005.08.020.

64. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research. 2018;47(D1):D886–D894. doi:10.1093/nar/gky1016.

65. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. Human Mutation. 2009;30(8):1237–1244. doi:10.1002/humu.21047.

66. Yip YL, Famiglietti M, Gos A, Duek PD, David FPA, Gateau A, et al. Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. Hum Mutat. 2008;29(3):361–366. doi:10.1002/humu.20671.

67. Chiti F, Dobson CM. Protein Misfolding, Functional Amyloid, and Human Disease. Annual Review of Biochemistry. 2006;75(1):333–366. doi:10.1146/annurev.biochem.75.101304.123901.

68. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982;157(1):105–32.