1 Contiguously-hydrophobic 2 sequences are functionally 3 significant throughout the human 4 exome

5 Ruchi Lohia^{1§}, Matthew E.B. Hansen^{2†}, Grace Brannigan^{1,3†*}

*For correspondence:

*grace.brannigan@rutgers.edu (GB)

[†]These authors contributed
equally to this work

Present address: [§]Stanley
Institute for Cognitive Genomics,
Cold Spring Harbor Laboratory,
USA

6 ¹Center for Computational and Integrative Biology, Rutgers University, Camden, NJ,
7 USA; ²Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA;
8 ³Department of Physics, Rutgers University, Camden, NJ, USA

9
10 **Abstract** Hydrophobic interactions have long been established as essential to stabilizing
11 structured proteins as well as drivers of aggregation, but the impact of hydrophobicity on the
12 functional significance of sequence variants has rarely been considered in a genome-wide
13 context. Here we test the role of hydrophobicity on functional impact using a set of 70,000
14 disease and non-disease associated single nucleotide polymorphisms (SNPs), using enrichment
15 of disease-association as an indicator of functionality. We find that functional impact is
16 uncorrelated with hydrophobicity of the SNP itself, and only weakly correlated with the average
17 local hydrophobicity, but is strongly correlated with both the size and minimum hydrophobicity
18 of the contiguous hydrophobic domain that contains the SNP. Disease-association is found to
19 vary by more than 6-fold as a function of contiguous hydrophobicity parameters, suggesting
20 utility as a prior for identifying causal variation. We further find signatures of differential selective
21 constraint on domain hydrophobicity, and that SNPs splitting a long hydrophobic region or
22 joining two short regions of contiguous hydrophobicity are particularly likely to be
23 disease-associated. Trends are preserved for both aggregating and non-aggregating proteins,
24 indicating that the role of contiguous hydrophobicity extends well beyond aggregation risk.

25 26 Statement of Significance

27 Proteins rely on the hydrophobic effect to maintain structure and interactions with the environ-
28 ment. Surprisingly, no signs that amino acid hydrophobicity influences natural selection have been
29 detected using modern genetic data. This may be because analyses that treat each amino acid sep-
30 arately do not reveal significant results, which we confirm here. However, because the hydrophobic
31 effect becomes more powerful as more hydrophobic molecules are introduced, we tested whether
32 unbroken stretches of hydrophobic amino acids are under selection. Using genetic variant data
33 from across the human genome, we found evidence that selection pressure increases continually
34 with the length of the unbroken hydrophobic sequence. These results could lead to improvements
35 in a wide range of genomic tools as well as insights into disease and protein evolutionary history.

36 Introduction

37 An ambitious task for human genetics is discovering the genetic basis for heritable traits and dis-
38 ease risks. In principle, genome-wide association studies (GWAS) can identify variants associated

with an effect, independent of the genomic or protein context of the variant (*Visscher et al., 2017*). In practice, 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 (*Cannon and Mohlke, 2018*). 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 (*Gallagher and Chen-Plotkin, 2018*).

All protein level methods for variant-to-function prediction rely on some form of residue characterization. In addition to physicochemical properties (*Stone, 2005; Niroula et al., 2015; López-Ferrando et al., 2017; Hecht et al., 2015; Popov et al., 2019*) these may include evolutionary conservation (*Ng and Henikoff, 2001; Thomas and Kejariwal, 2004; Stone, 2005; Capriotti et al., 2006; Choi et al., 2012; Hecht et al., 2015; Niroula et al., 2015*) and structural propensities (*Capriotti et al., 2004, 2005; Parthiban et al., 2006; Wainreb et al., 2010; Popov et al., 2019; Ittisoponpisan et al., 2019; Iqbal et al., 2020*). Such methods may also rely on known protein structures in order to incorporate properties such as local secondary structure and solvent accessibility (*Iqbal et al., 2020*). 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 known proteins (*Rose et al., 2016; Mir et al., 2017*). Fewer than 35% of human protein coding genes have structures deposited in the protein data bank (*Prlić et al., 2016*). With few exceptions, therefore, these methods have not been applied on a genome-wide scale.

Most functional proteins are intrinsically modular. For example, 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. In the absence of structural information, bioinformatics “variant-to-function” prediction methods that use sequence context typically neglect modularity, by defining the local sequence using a window of constant length, centered around the mutation (*Teng et al., 2010; Capriotti et al., 2006; Capriotti and Fariselli, 2017; Hecht et al., 2015*). This definition of sequence context can weaken predictive power, particularly for SNPs near the boundary between adjacent protein modules with distinct functional roles.

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 are both essential for many critical biological pathways (*Ward et al., 2004; Dyson and Wright, 2005; Uversky, 2013, 2019; Panchenko and Babu, 2015*) and sensitive to sequence. (*Weinreb et al., 1996; Uversky et al., 2012; Cuanalo-Contreras et al., 2013; Patel et al., 2015; Uversky, 2015; Tovo-Rodrigues et al., 2016*).

Although not all functional proteins have well-defined structures, we showed in a previous study (*Lohia et al., 2019*) that a long intrinsically disordered protein can retain modularity of interactions. Fully-atomistic molecular dynamics simulations of a 92-residue disordered protein, run using enhanced sampling, revealed a soft network of tertiary contacts between contiguous stretches of hydrophobic residues. In order to distinguish these stretches from any other more traditional segment or domain definition, we follow terminology more common to polymer physics and call these stretches “blobs”. Blobs may contain secondary structure elements, but are not required to do so.

Here we demonstrate that hydrophobic blobs can be used as generic predictors of functional modules, by testing for enrichment of functional effects throughout the proteome. Our hypothesis was motivated by the cooperativity of the hydrophobic effect (*Jiang et al., 2017*) and the recognition that hydrophobic residues are likely to be buried (*Lins et al., 2003*) and contain a high density of interactions. Any curve drawn through such a cluster, including a curve representing the peptide chain, will contain contiguous hydrophobic residues (Figure 1).

In the present study, we used a large database of protein sequences to determine the extent

to which hydrophobicity-based sequence segmentation captures functional modules. Our test involves three steps: 1) segmentation of the protein into the blobs ("blobulation"), 2) characterization of the blobs based on various physicochemical properties and 3) testing whether SNPs with "known" functional impact are enriched or depleted genome-wide in blobs with particular properties. We test for enrichment of hydrophobic blobs among disease-associated SNPs, repeating the calculation with a range of hydrophobicity thresholds and stratifying according to hydrophobic blob length. This process reveals a striking dependence that is muted when a standard moving window approach is used instead, and is completely lost when the sequence context is neglected. We test for consistency against population frequency data, finding that disease-associated SNPs in extremely hydrophobic blobs are under greater purifying selection than those in less hydrophobic blobs. We further test the functional effects of mutations that split or merge blobs or change other categorical properties. In doing so, we find that blobulation yields a meaningful protein topology that is useful for sequence characterization beyond hydrophobicity.

Results

Regions of contiguous hydrophobicity are enriched for disease-causing SNPs

Figure 1. Cartoon representation of the blobulation approach. Left: A given protein chain consists of amino acids that are either hydrophobic (blue ovals) or not hydrophobic (orange fans). Clusters of hydrophobic residues (gray shading) are likely to be buried and have a high number of interacting residues. Any curve drawn through such a cluster, including the curve representing the peptide chain, will contain contiguous hydrophobic residues. Center: The blobulation algorithm calculates the smoothed hydrophobicity for each residue, and then identifies stretches of at least $L_{\min}$ residues above the hydrophobicity threshold (black line), where $L_{\min} = 4$ unless otherwise noted. This process defines the boundaries between blobs, and also provides information about the blob hydrophobicity class: blobs of contiguous hydrophobic residues are "h" blobs (at least $L_{\min}$ residues), while remaining stretches are either p "blobs" (at least $L_{\min}$ residues) and "s blobs" (fewer than $L_{\min}$ residues). Right: Blobs mapped onto the peptide chain; the peptide backbone is colored according to blob hydrophobicity class and black circles mark the boundary between blobs. In this illustration there are 4 h-blobs, 2 p-blobs, and 1 s-blob.

In order to test whether the hydrophobicity of a reference allele was correlated with its functional impact, we calculated the enrichment of dSNPs relative to nSNPs as a function of hydrophobicity of the reference allele. As shown in Figure 2b, we did not detect any significant correlation (Pearson's $r=0.02$, $n=17$). Hydrophobicity is a highly cooperative phenomenon (Jiang et al., 2017), and we hypothesized that the hydrophobicity of the local sequence could be critical for determining the functional effects of the dSNP. We considered two methods for determining local sequence context: our own "blobulation" approach and a fixed-width moving window.

Blobulation (Figure 1) is a particular systematic approach for identifying modular components in the protein sequence, which we previously introduced (Lohia et al., 2019) for analysis of a specific protein. Here we have generalized and extended the approach by varying two previously-fixed parameters (minimum blob length and hydrophobicity threshold) that are required to define blob edges, as described in Methods. The degree of enrichment for disease-association is expected to be dependent upon both of these parameters, and we ran enrichment tests for a wide range of minimum blob lengths and hydrophobicity threshold values. Unless otherwise noted, dSNPs are

119 tested for enrichment relative to the expectation set by nSNPs . For example, the phrase “dSNPs are
120 enriched in X blobs” means that dSNPs are found at a higher rate in blobs of type X than are nSNPs .

121 We found a consistent trend relating hydrophobicity of the local blob and enrichment of disease-
122 associated SNPs (Figure 2 C). dSNPs are depleted in weakly hydrophobic blobs, neutral for mod-
123 erately hydrophobic blobs, and become more enriched as the blob gets longer and/or satisfies
124 a stricter hydrophobicity threshold. The trend is steadily monotonic, which supports the hypoth-
125 esis that hydrophobic blobs constitute identifiable functional elements. The exception is at the
126 plot boundary: surprisingly, blobs that satisfied the very strictest criteria were depleted in dSNPs .
127 In the next section, we consider two potential reasons for this depletion: either dSNPs in these
128 blobs are so deleterious that they are selected out of the population, or nSNPs in these blobs are
129 functional as expected, but not deleterious. In addition to a consistent trend, the analysis returns a
130 large spread in enrichment/depletion values: 3.2% of the bins in Figure 2c have a significant enrich-
131 ment of less than 0.5, while 13% have a significant enrichment of greater than 1.5, and 3.5% have a
132 significant enrichment greater than 3 (Data set 1). This result indicates that hydrophobicity-based
133 sequence segmentation could be particularly useful for assessing the riskiness of SNPs located in
134 long and very hydrophobic sequences.

135 In order to compare the effectiveness of blobulation to a moving window approach, we ran
136 analogous enrichment tests for sliding windows across all protein sequences. Two analogous pa-
137 rameters (window length and mean hydrophobicity threshold) are used for defining hydrophobic
138 regions in the moving window approach, with the resulting enrichment for a range of these values
139 also shown in Fig 2d. Compared to the blobulation approach, the enrichment detected using the
140 moving window approach is much less sensitive to segment length for windows beyond 6 residues
141 (consider the tilted color bands for the blobulation approach vs the horizontal color bands for mov-
142 ing windows in Figure 2c-d). The significant depletion in dSNPs at the most hydrophobic edge is
143 also preserved in the moving window analysis.

144 While there is no “standard” window size, most SNP prediction programs use a window size in
145 the range of 1-21 residues (*Teng et al., 2010; Capriotti et al., 2006; Capriotti and Fariselli, 2017;*
146 *Hecht et al., 2015*). The window size is chosen to balance concerns that small window sizes may
147 not accurately capture the “local” sequence (*Chen et al., 2006; Schlessinger and Rost, 2005; Sander*
148 *et al., 2006*) whereas long window sizes can increase signal to noise ratio (*Park et al., 2007*). In
149 either case, the use of a fixed-width window neglects the inherent dispersion in the size of protein
150 modules (Figure 2 a). The enrichment insensitivity to window size was thus surprising, but the
151 expected noise was likely lost due to averaging across the proteome. As shown in Fig 2e, the
152 overall fraction of hydrophobic segments for mild or moderate hydrophobicity thresholds was also
153 insensitive to window size, confirming that the data set we use here is sufficiently large to average
154 out large window noise. For a single SNP, considering too many residues far from the SNP will still
155 contribute significant noise to the prediction.

156 Overall, use of the moving window approach yields two-fold enrichment for only a few, scat-
157 tered parameter values with very few qualifying sequences. In contrast, blobulation yielded greater
158 than two-fold enrichment for a well-defined set of parameters (hydrophobicity threshold > 0.4 and
159 minimum blob size > 15). This observation supports our hypothesis that blobulation provides a
160 more meaningful and less noisy approach to protein segmentation than use of a fixed-length mov-
161 ing window.

162 **Population frequency signatures of differential selection on hydrophobic blobs**

163 If hydrophobic blobs capture functional segments of coding regions, then we may expect to see
164 signatures of differential selective constraint with varying blob properties. We probe this by exam-
165 ining the genetic diversity of SNPs, stratified by their surrounding blob hydrophobicity. A bench-
166 mark measure of genetic diversity is the expected heterozygosity $\Theta \equiv 2\nu(1 - \nu)$, where ν is the
167 frequency of the coded allele. Sequences under more functional constraint, like exons in essential
168 proteins, experience purifying selection (removal of almost all new functional alleles), which low-

Figure 2. Effect of segmentation approach, length, and hydrophobicity threshold on calculated enrichment of dSNPs in hydrophobic segments. a) Illustration of three measures of SNP hydrophobicity (residue, contiguous, and average) for the amino acid shown in purple, found within a hypothetical peptide chain where hydrophobic residues are blue ovals and non-hydrophobic residues are orange fans. The local sequence determined by each segmentation method is shown in black fill. b) Enrichment of dSNPs relative to nSNPs as a function of hydrophobicity of the reference allele, with line of best fit. No trend or significant correlation is observed (Pearson's $r=0.02$, $p=0.94$, $n=17$). c-d) Enrichment of dSNPs relative to nSNPs for hydrophobic segments for varying hydrophobicity and length thresholds using a blobulation approach (c) and moving window approach (d). Bins with no data are shown in gray. e-f) The number of nSNPs in hydrophobic segments per bin using the blobulation approach (e) and moving window approach (f).

ers Θ compared to regions under little or no constraint (see, e.g. *Dickerson (1971); Charlesworth et al. (1993); Sunyaev et al. (2000); Siepel et al. (2005); Somel et al. (2013)*). Conversely, balancing selection (maintenance of multiple functional alleles) causes increased genetic diversity over a genomic region (see, e.g. *Llaurens et al. (2017)*). Population substructure can also increase the genetic diversity, but such effects would occur genome-wide and would not be correlated with blob hydrophobicity class.

For population frequency data we use gnomAD v3 SNP frequency data for $N \sim 32,000$ non-Finnish Europeans *Karczewski et al. (2021)*. The average heterozygosity of non disease-associated SNPs over this cohort is $\langle \Theta \rangle_{\text{nSNP}} = 0.14$, while that of disease-associated SNPs is nearly 50 times lower, $\langle \Theta \rangle_{\text{dSNP}} = 0.0029$. Figure 3 shows the mean heterozygosity of nSNPs and dSNPs binned by the maximal blob hydrophobicity (H_{max}) for each SNP (see Methods). The null background is generated by random permutation of the heterozygosity assigned to each SNP, corresponding to the null hypothesis that there is no correlation between SNP heterozygosity and blob properties. Both nSNPs and dSNPs show a significant departure (outside the 1st – 99th null percentiles) in mean heterozygosity from null expectations for the largest H_{max} bin, $H_{\text{max}} \geq 0.75$. For nSNPs, the heterozygosity is larger than expected in the highest H_{max} bin ($> 99^{\text{th}}$ null percentile), consistent with relaxed purifying selection or increased balancing selection. In contrast, for dSNPs the heterozygosity in the largest H_{max} bin is smaller than expected under the null ($< 1^{\text{st}}$ null percentile). Given that disease risk alleles are almost always “new” deleterious mutations, we surmise that this decrease in Θ is due to greater purifying selection acting to remove the risk allele. In other words,

Figure 3. Expected heterozygosity of SNPs in Europeans as a function of blob hydrophobicity The average expected heterozygosity θ for SNPs in h-blobs as a function of the maximum SNP hydrophobicity $H_{\max}$ (binned in bin widths of $\Delta H_{\max} = 0.25$) is shown by the black line, with standard errors in the mean indicated. Only SNPs where at least one of two alleles is in an annotated h-blob are considered. Frequencies are taken from the gnomAD cohort of non-Finnish Europeans. The brown line is the average heterozygosity for all SNPs with $H_{\max} < 0.5$ (the line is solid for $H_{\max} < 0.5$ and dashed for $H_{\max} > 0.5$). The background null distribution generated from randomized permutations are shown in brown, with the null percentiles as shown in the legend on the right. Panels a) and b) show the results for nSNPs and dSNPs, respectively.

the disease-associated alleles in highly hydrophobic blobs appear to be more deleterious than disease-associated alleles in lower hydrophobicity blobs.

The results here may explain the non-monotonic behavior of the dSNP-to-nSNP enrichment seen in Figure 2c and d for both the blobulation and moving windows calculations. The thin band of “red” over the bins with the highest hydrophobic threshold that still contain data indicates a depletion in the number of dSNPs relative to nSNPs. Before examining the population frequency data, it was not clear if the lack of dSNPs relative to nSNPs is due to increased selection against dSNPs in those blobs, increased balancing selection for nSNPs in those blobs, or a mixture of both. In this section we have shown that the heterozygosity of dSNPs decreases with higher blob hydrophobicity while the heterozygosity of nSNPs rises with blob hydrophobicity. This suggests that the depletion observed in Figure 2c and d in the high hydrophobicity regions is caused by both increased selection against disease-associated alleles and increased selection for the maintenance of functional non disease-associated alleles.

We tested whether the trends observed in Figure 3 are specific to non-Finnish Europeans by performing the same analysis on the gnomAD East Asian frequency data set of approximately ~ 1,500 individuals. Although we do not observe the same level of statistical significance, we do find the same trends as in the European cohort: for nSNPs there is increasing heterozygosity with increasing $H_{\max}$ and for dSNPs the heterozygosity decreases with increasing $H_{\max}$ (Figure S1). The stratification of genetic diversity with blob properties appears to be a feature shared across diverse human populations.

In order to provide some context for the proteins containing the blobs in the $H_{\max} \geq 0.75$ bin we use gene-ontology and pathway enrichment tests. The nSNPs found in blobs with $H_{\max} \geq 0.75$ are distributed across a subset of proteins ($n = 635$), which we test against the background of all proteins containing nSNPs ($N = 10,406$). For the dSNPs, there are $n = 179$ proteins containing the $H_{\max} \geq 0.75$ SNPs and a total background of $N = 1803$ proteins containing a dSNP. Both the $H_{\max} \geq 0.75$ nSNP and dSNP proteins are most enriched for being located in or on a membrane (Data sets 2 and 3), as expected for proteins containing highly hydrophobic blobs. In terms of molecular function and biological process, however, the nSNPs and dSNPs differ in their enriched ontologies. The $H_{\max} \geq 0.75$ nSNPs are enriched for olfactory processes, chemical sensing, and

218 signal transduction. This accords with previous observations that olfactory-related genes exhibit
 219 signs of relaxed purifying selection relative to other genic regions *Pierron et al. (2013)*. This analysis
 220 has essentially re-identified this same feature of greater diversity in chemical sensing pathways,
 221 found here because the increased genetic diversity is specifically present in highly hydrophobic
 222 subdomains of those proteins. In contrast, the $H_{\max} \geq 0.75$ dSNPs are enriched for cation and
 223 small molecule trans-membrane transport proteins. These SNPs reside in proteins enriched for
 224 critical membrane bound transporters, consistent with the signal of higher functional constraint.

225 **dSNP enrichment emerges in shorter and more weakly-hydrophobic blobs for ag-** 226 **gregating proteins**

Figure 4. Blob hydrophobicity for SNPs in aggregating proteins. a) Fraction of nSNPs or dSNPs that are found in either blob type (a) or charge class(b) in non-aggregating proteins and known-aggregating proteins. c) is the fraction of charge class in each blob type in all proteins. The distribution in a, b and c 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 2b. Error bars represent one standard error for multinomial distributed data. d) is same as Fig 2b but for non aggregating proteins and aggregating proteins. Enrichment of dSNPs in h-blob at fixed cutoff of 0.4 (e) or fixed blob length of 6.

227 Side-chain hydrophobicity plays a well-established role in diseases involving aggregating pro-
 228 teins. (*Conchillo-Solé et al., 2007; Fang et al., 2013; Tartaglia and Vendruscolo, 2008*). In order
 229 to test whether the trends observed in Fig 2b are amplified in aggregating proteins, we separated
 230 aggregating proteins from the primary data set. Proteins involved in the formation of extracellu-
 231 lar amyloid deposits or intracellular inclusions with amyloid-like characteristics are included in the

232 "Aggregating proteins" (28 proteins, 124 nSNPs, 330 dSNPs) subset, while all remaining proteins
233 are labelled as "Non-aggregating proteins".

234 Figure 4a shows the blob-type distribution of nSNPs and dSNPs, for both aggregating and non-
235 aggregating proteins, using a mild hydrophobicity cutoff of 0.4. For dSNPs, the blob distributions
236 are insensitive to aggregating-tendencies of the protein: a given dSNP in an aggregating protein is
237 as likely to be found in an h-blob as if it were in a non-aggregating protein. We do still detect a small
238 difference in enrichment because nSNPs in aggregating proteins are found slightly less frequently
239 in h-blobs vs nSNPs in non-aggregating proteins.

240 Stratifying the enrichment by hydrophobicity cutoff and blob length (as in Fig 2b of the main
241 text) reveals a more informative quantitative shift. Figure 4d compares the data from Fig 2b) with
242 the same analysis for aggregating proteins. The highly-enriched (blue) band is shifted toward the
243 origin for aggregating proteins, indicating that sensitivity to mutation is found in shorter and more
244 weakly hydrophobic blobs. This shift is also visible in transects of the data along one row and one
245 column, shown in Figure 4e and f. These results are consistent with previous observations that the
246 in vitro aggregation rate of unstructured polypeptide chains is proportional to hydrophobicity and
247 inversely proportional to net charge (DuBay et al., 2004).

248 To test the significance of charge, blobs were also assigned charge class (Figure 4b), which is the
249 predicted globular phase according to Das and Pappu (Das and Pappu, 2013) and calculated using
250 the fraction of positive and negative charges. Possible values of the blob charge class are 1 (Weak
251 polyampholyte), 2 (Janus or boundary region), 3 (Strong polyampholyte), 4 (Negatively-charged
252 strong polyelectrolyte), and 5 (Positively-charged strong polyelectrolyte). Sequences in class 1 have
253 a low fraction of positively charged residues, and nearly all structured proteins fall in class 1. In
254 contrast to structured proteins, IDPs can be found in all five Das-Pappu phases, particularly classes
255 3, 4, and 5. The blob hydrophobicity class and charge class are fundamentally correlated; while
256 blob charge class does not explicitly consider hydrophobicity, increasing the number of charged
257 residues will reduce the average hydrophobicity of a blob. The extent of this correlation is shown
258 in Figure 4C, which breaks down the fraction of h and p blobs that fall in each Das-Pappu charge
259 class. As expected, most h blobs (90%) fall in class 1 (weak polyampholyte), followed by 9% in class
260 2 (Janus). The p blobs are more evenly distributed across classes, with the highest fraction (42%)
261 classified as strong polyelectrolytes.

262 Since there are five charge classes and only three hydrophobicity classes, we hypothesized
263 that blob charge class would be more precise than blob hydrophobicity class for identification of
264 functional protein segments. Instead, we found that the maximum charge class enrichment (1.09
265 fold for weak polyampholyte blobs) is comparable or even slightly less than the enrichment found
266 for hydrophobicity class (1.15 fold). This result suggests that even short contiguous stretches of
267 weak hydrophobicity are as effective as charge distribution for predicting functional impact. In
268 the context of aggregating proteins, classification by charge yielded similar patterns as classifi-
269 cation by hydrophobicity: dSNPs have equivalent distributions regardless of protein aggregation
270 tendency, while nSNPs in aggregating proteins are significantly less likely to be found in globular,
271 weak-polyampholyte blobs (Class 1) than those in non-aggregating proteins.

272 **Disease-associated SNPs are enriched for mutations that change local blob charac-** 273 **teristics and overall protein blob topology**

274 The blobulation process yields a series of h blobs, connected by p and s linkers, which we term
275 the "blobular topology". Such a topology is analogous to the classic protein topology of secondary
276 structure elements. A SNP can alter the blobular topology by moving a short stretch of contiguous
277 residues above or below the minimum blob size, either forming a new small h-blob or dissolving an
278 existing small h-blob respectively. In the background case we expect to see more formation than
279 dissolution, since the blob frequency decreases with length (Figure 2 e), and there are more blobs
280 just below the minimum length than just above it. Figure 5A confirms this expectation for nSNPs.
281 A SNP can also split a long h-blob by interrupting a long contiguous hydrophobic sequence, or

Figure 5. Distribution of SNPs that change blob properties a) Fraction of nSNPs or dSNPs that change the blobular topology by either forming or dissolving an h-blob or by splitting one h-blob or merging two h-blobs. b) Fractions of nSNPs or dSNPs that induce a change in hydrophobicity class or charge class. c) The enrichment of dSNPs relative to nSNPs that induce a specific blobular topology transition between the reference allele (x-axis) and the alternative allele (y-axis). This is shown for two blob properties, hydrophobicity class (left) and charge class (right). d) The overall proportion of nSNPs that induce each of the blobular topology transitions shown in c). Fewer than 1.5% of SNPs involve charge classes 4 or 5, and these data are not shown for readability. Significant enrichment or depletion in dSNPs is annotated with one star ($p < 10^{-3}$) or two stars ($p < 10^{-10}$). Errors bars in (a) and (b) represent one standard error for multinomial (a) or binomial (b) distributed data. In all cases blobulation is calculated using $H^* = 0.4$ and $L_{\min} = 4$.

merge two smaller h-blobs into one long h-blob by removing such an interruption. Using the mild hydrophobicity threshold of 0.4 and minimum blob length of 4 residues, fewer than 15% of SNPs overall cause a topological change. As shown in Figure 5C, however, certain topological changes are strongly enriched in dSNPs. While the difference between the fraction of nSNPs and dSNPs that form h-blobs is not significant ($p > 10^{-3}$), dSNPs are significantly more likely to dissolve h-blobs ($p < 10^{-21}$), split h-blobs ($p < 10^{-80}$), and merge h-blobs ($p < 10^{-36}$). The magnitude of enrichment (2.1-fold) is greatest for SNPs that split longer h-blobs into two shorter ones and is only moderately weaker for the reverse merging of two h-blobs (1.7-fold). These results are consistent with the functional sensitivity of long blobs shown in Figure 2c.

Regardless of overall topological changes, SNPs may change the characteristics of their local blob. Such changes may affect blob topology (as in Figure 5a), or simply shift blob boundaries (as in Figure 5b). We now broaden the analysis to include changes in blob characteristics that do

not necessarily affect the blob boundaries. We observe that about 17% of the nSNPs and 22% of dSNPs introduce a blob-hydrophobicity class change, yielding a 1.3 fold enrichment for blob hydrophobicity class changes (Fig 5b). Fig 5c compares the frequency of possible blob hydrophobicity class transitions. Mutations involving h $\rightarrow$ p blob transitions yield the maximum enrichment (1.5 fold) among dSNPs. In contrast, dSNPs are only 1.1 fold enriched in the reverse transition (p $\rightarrow$ h), and SNPs that remain in p blobs for both the ancestral and derived allele in p blobs are depleted among disease-associated SNPs.

Similarly, the frequency of SNP-induced changes in blob charge class is shown in Fig 5c, for transitions involving blobs in categories 1, 2, and 3. Transitions involving categories 4 or 5 (positively or negatively-charged strong poly electrolyte) represented fewer than 1.5 % 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 the local blob charge class. so 1 $\rightarrow$ 3 $>$ 1 $\rightarrow$ 2 and 3 $\rightarrow$ 1 $>$ 3 $\rightarrow$ 2. Collectively, these results are consistent with the increased mutational sensitivity of hydrophobic (and typically buried) blobs that is shown in Figure 2, while also emphasizing that mutations that *change* charge class are particularly likely to be causal. For instance, charge reversal of a charged residue could have particularly strong functional effects. While these effects might be amplified if the residue was located in an h-blob (and thus interacting with other protein residues), the charge reversal itself would not affect the blob hydrophobicity class.

Discussion

In the present work, we have presented the “blobulation” scheme for identifying interaction-rich protein regions from sequence, and tested its use in detecting functional modules across the proteome. We find that enrichment of disease-associated mutations in hydrophobic blobs increases with the strictness of the hydrophobic blob criteria, with greater than 4-fold enrichment for disease-association in the longest, most hydrophobic blobs. The range and resolution of varying enrichment is strongly damped in the *status quo* fixed-length moving window approach. Stratifying SNPs by their surrounding blob properties reveals genome-wide differences in blob genetic diversity, demonstrating pervasive differential selection on at least some classes of blobs.

We also find that disease-associated mutations are significantly more likely than non-disease associated mutations to change the topology of the sequence. This suggests that blobulation provides a meaningful topology that can be used as a framework for sequence analysis, and which requires only the protein amino acid sequence and two parameters (minimum blob length and hydrophobicity threshold). Once blobs are identified, they can be characterized on any property of interest. As an example, in the present work, we also find that disease-associated mutations are moderately enriched for mutations that cause transitions in blob hydrophobicity class (up to 1.5 fold) and strongly enriched for mutations that cause certain transitions in blob charge class (1.7 fold).

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 (Conchillo-Solé *et al.*, 2007), ProA (Fang *et al.*, 2013), and Zyggregator (Tartaglia and Vendruscolo, 2008). We do find that functional sensitivity in aggregating proteins follows similar trends but emerges in shorter blobs satisfying weaker hydrophobicity criteria. The difference between aggregating and generic proteins is thus quantitative, not qualitative, and demonstrates that hydrophobic interactions occur on a useful continuum.

Structural data implicitly provides insight into the segmentation of a protein sequence, but we show here that segmentation may be meaningfully estimated from the sequence alone. We have proposed blobulation as a generalizeable approach for segmentation, although 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 (Yang et al., 2018; Zhang et al., 2018; Wang et al., 2017; Ma et al., 2018). 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.

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 (Claussnitzer et al., 2020). SNP-prediction methods often test residue properties for relevance by machine learning algorithms, which 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 both performed hypothesis-driven tests for enrichment of certain features in putatively causal SNPs and provided a two-dimensional table (Dataset 1) of wide-ranging enrichments as a function of blob properties, which can be used as inputs in future predictions.

Methods and Materials

SNP data

The SNP data we use is from the UniProtKB literature curated list of missense variants (www.uniprot.org/docs/humsavar, obtained on 17-Jun-2020) (Yip et al., 2008). Variants are annotated using the American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) terminology (Richards et al., 2015). dSNPs are those annotated as “likely pathogenic or pathogenic” ($N = 30, 227$), and nSNPs are those annotated as “likely benign or benign” ($N = 39, 448$).

Blobulation

Blobulation refers to the partitioning of a given amino acid sequence (e.g. a protein isoform) into three kinds of segments based on the local hydrophobicity. The segments are referred to as “blobs”, and they are assigned a primary type known as the “hydrophobicity class”. The three blob hydrophobicity classes are hydrophobic (“h-blob”), non-hydrophobic or polar (“p-blob”), and short non-hydrophobic spacers (“s-blobs”). We use a simple partitioning approach, described below, that is based on three parameters: the h-blob hydrophobicity threshold H^* , the minimum length for hydrophobic segments to be labeled as an h-blob $L_{\min}^h$, and the threshold length separating p-blobs from s-blobs, $L_{\min}^p$.

Every amino acid i is assigned a mean hydrophobicity H_i , defined as the average Kyte-Doolittle (Kyte and Doolittle, 1982) score with a window size of three residues, scaled to fit between 0 and 1. Given the hydrophobicity threshold H^* , we first identify all contiguous stretches where $H_i \geq H^*$ for all i in the sub-sequence and where the sub-sequence length x is $x \geq L_{\min}^h$; these are classified as h-blobs. The sub-sequence between a pair of h-blobs, or between an h-blob and the end of the amino acid sequence (for sub-sequences at an edge), is classified as a p-blob if the length $x \geq L_{\min}^p$, and is classified as an s-blob if $x < L_{\min}^p$. In all analyses used in this work, we set $L_{\min}^p = L_{\min}^h \equiv L_{\min}$, so that there are effectively only two blobulation parameters, H^* and $L_{\min}$.

Consider a given protein amino acid sequence that contains a missense SNP at residue i . For a missense SNP, the reference allele and the alternative allele correspond to two different amino acids. We index the two amino acids for this polymorphism by $\alpha \in \{R, A\}$, where R is the amino acid corresponding to the reference allele and A for the alternative allele. Upon blobulation of this sequence with amino acid α at i , the residue i will be contained within a blob type $\beta \in \{h, p, s\}$ with some length x . We define the binary blobulation query function $B_{\alpha, \beta, x}(i)$

$$B_{\alpha, \beta, x}(i) = \begin{cases} 1, & \text{if the residue } i \text{ with allele } \alpha \text{ is in an } \beta \text{ blob of length } x, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Importantly, $B_{\alpha,\beta,x}(i)$ is dependent upon the hydrophobicity threshold H^* (as are all quantities that depend on $B_{\alpha,\beta,x}(i)$). Furthermore, because h-blobs are detected first, p-blobs are detected from the remaining residues, and s-blobs are assigned last, $B_{\alpha,p,x}(i)$ will be dependent upon the minimum h-blob length, and $B_{\alpha,s,x}(i)$ will be dependent upon the minimum h and p-blob lengths. Unless stated otherwise, blobulation is based on the reference allele for all SNPs and we suppress the explicit α notation.

Das-Pappu charge class

The blob charge class is a secondary blob property representing the Das-Pappu phase (*Das and Pappu, 2013*), which is determined using the fraction of positively and negatively charged residues. Blobulation does not rely on charge class, but any blob may be assigned a charge class following blobulation. 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).

dSNP enrichment

Let $f_{\beta,x}$ and $g_{\beta,x}$ denote the fraction of nSNPs and dSNPs, respectively, which are found in β -type blobs of length x :

$$f_{\beta,x} = \frac{1}{N_{\text{tot}}} \sum_{i \in \mathcal{N}} B_{\beta,x}(i) \quad (2)$$

$$g_{\beta,x} = \frac{1}{D_{\text{tot}}} \sum_{i \in \mathcal{D}} B_{\beta,x}(i) \quad (3)$$

where $\mathcal{N}$ is the set of nSNPs with size $N_{\text{tot}} \equiv \|\mathcal{N}\|$ and $\mathcal{D}$ is the set of dSNPs with size $D_{\text{tot}} \equiv \|\mathcal{D}\|$. Enrichment of dSNPs relative to nSNPs for β -blobs of length x is then

$$v_{\beta}(x) = \frac{g_{\beta,x}}{f_{\beta,x}}. \quad (4)$$

For some comparisons we consider the enrichment over all blob lengths above some minimum threshold $x \geq x^*$: $F_{\beta} \equiv \sum_{x \geq x^*} f_{\beta,x}$ and $G_{\beta} \equiv \sum_{x \geq x^*} g_{\beta,x}$ for nSNPs and dSNPs, respectively. The corresponding minimum-length enrichment is $Y_{\beta} = G_{\beta}/F_{\beta}$. For the heat-map in Figure 2 we consider the enrichment over SNPs at a given hydrophobicity threshold H^* for lengths within a length bin. Within a length bin $[x_0, x_1]$, the fractions are summed over all x within that bin: $F_{\beta}^{\text{bin}} \equiv \sum_{x=x_0}^{x_1} f_{\beta,x}$ for nSNPs and similarly for dSNPs. The enrichment is given by the ratio of these bin fractions, $Y_{\beta}^{\text{bin}} = G_{\beta}^{\text{bin}}/F_{\beta}^{\text{bin}}$.

Enrichment values are tested for statistically significant deviations from the null expectation of $v = 1$ using a Binomial distribution for $N_{\text{trials}} = D_{\text{tot}}$ Bernoulli trials with probability of success $p = f_{\beta,x}$ per trial, where the expected number of “successes” is $D_{\text{tot}} f_{\beta,x}$. The minimum-length enrichment Y , and enrichment within length bins, are tested similarly. Reported p-values correspond to the two-sided test, and any enrichment or depletion in dSNPs is termed “significant” for p-value $< 10^{-3}$.

Blob transitions induced by a SNP

For each protein sequence and each SNP i in the given sequence, the blobulation algorithm is performed separately for both the reference and the alternative allele. If the protein contains multiple SNPs, then the reference allele is used for all residues except i , i.e. we do not consider multi-SNP *cis* effects on blobulation. In order to test the statistical significance of the observed transition rates for dSNPs, we compare the observed rates to those for nSNPs. Specifically, suppose the observed fraction of nSNPs that induce a transition from blob class β_R for the reference allele to blob class β_A for the alternative allele is $f(\beta_R, \beta_A)$, and that the fraction of dSNPs that induce such a transition is $g(\beta_R, \beta_A)$. Then the number of Bernoulli trials is $N_{\text{trials}} = D_{\text{tot}}$ with probability of success $p = f(\beta_R, \beta_A)$, where the expected number of “successes” is $D_{\text{tot}} f(\beta_R, \beta_A)$ and the observed number of successes is $D_{\text{tot}} g(\beta_R, \beta_A)$. Reported p-values correspond to the two-sided test, and any enrichment or depletion in dSNPs is termed “significant” for p-value $< 10^{-3}$.

Fixed-length moving windows

Calculations involving fixed-length moving windows apply a threshold test to the mean hydrophobicity across the window, rather than testing each residue. The mean hydrophobicity of a fixed-length moving window of length x (for odd x) centered around residue i is:

$$\bar{H}_{x,i} = \frac{1}{x} \sum_{j=-(x-1)/2}^{(x-1)/2} H_{i+j} \quad (5)$$

and the associated query function for the moving-window equivalent of h-blobs ($\beta = h$) with the reference allele at i ($\alpha = R$) is:

$$W_x(i) = \begin{cases} 1, & \text{if } \bar{H}_{i,x} \geq H^* \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

The query function W is used in place of B for Equations 2 and 3. All subsequent equations remain unchanged.

Population frequency data

Frequency data is from the gnomAD v3 genomes dataset of variants for the non-Finnish European cohort: gnomad.genomes.r3.0.sites.chr*.vcf.bgz, accessed July 24, 2020, using the "INFO/AF_nfe_male" and "INFO/AF_nfe_female" tags. This cohort contains 32,299 individuals, providing allele frequency data as low as $\sim 1/60,000 \sim 0.002\%$. Variants with dbSNP identifications ("rsids") were intersected with the UniProt SNP data. There are 36,025 SNPs in common (in 10,565 genes), composed of 29,653 nSNPs (in 10,143 genes) and 6,372 dSNPs (in 1,594 genes). For comparison we also analyzed the gnomAD v3 East Asian cohort, using the "INFO/AF_eas_male" and "INFO/AF_eas_female" tags. This cohort contains 1,567 individuals.

Expected heterozygosity of SNPs in h-blobs

We test for the impact of the blob hydrophobicity on SNP heterozygosity by comparing the average expected heterozygosity to the null distribution in different hydrophobicity bins, described as follows. We first partition SNPs by the maximum blob hydrophobicity threshold for which at least one of the alleles remains in an h-blob with length $\geq L_{\min}$, defined as $H_{\max}$. SNPs where neither allele reside in an h-blob with this minimum length are discarded from further analysis. The expected heterozygosity of a SNP with frequency ν (for either the reference or alternative allele) is defined as $\Theta = 2\nu(1 - \nu)$. Define the blob length at $H_{\max}$ as L^* . We bin SNPs into four $H_{\max}$ bins of width $\Delta H_{\max} = 0.25$ and compute the mean SNP expected heterozygosity within each bin. The null distribution for each bin is computed based on $R = 5,000$ random permutations of the SNP heterozygosity among the input SNP set. The preceding procedure is tabulated for nSNPs and dSNPs separately.

Identification of aggregating proteins

There are 28 proteins within the dataset that are annotated as involved in formation of extracellular amyloid deposits or intracellular inclusions with amyloid-like characteristics: 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. For analyses involving aggregation, the uniprot dataset was divided into the SNPs within these 28 proteins ("Aggregating Proteins") and all other proteins ("non-Aggregating Proteins").

Gene pathway and ontology enrichment tests

The gene-ontology enrichment is performed using the g:Profiler web service [Raudvere et al. \(2019\)](#). We use the g:GOST functional annotation enrichment tool. Statistical p-values are adjusted for multiple-testing and ontology overlap using the g:Profiler algorithm "g:SCS" on a user significance

threshold of 0.05. We use custom backgrounds over annotated genes only. The background used for nSNPs is all proteins containing an nSNP and the background for dSNPs is all proteins containing a dSNP. The ontology databases tested for enrichment are the g:Profiler databases as of 2019: GO Molecular Function (GO MF), GO Biological Process (GO BP), GO Cellular Component (GO CC), Kyoto Encyclopedia of Genes and Genomes pathways (KEGG), Reactome pathways (REAC), WikiPathways (WP), TRANSFAC (TF), miRTarBase (MIRNA), the Human Protein Atlas (HPA), the Comprehensive Resource of Mammalian protein complexes (CORUM), and the human phenotype ontology (HP).

Statistical Information

Throughout this paper, enrichment of dSNPs relative to nSNPs is tested using the Binomial distribution for single category tests, and the multinomial distribution for multi-category tests. In general, the number of tests is the number of dSNPs, with the per-test probability of “success” given by the nSNP rate. The probability of having a number of successes at least as extreme as the observed number of dSNPs is the reported (two-sided) p-value. See the above subsections for detailed numbers for each of the enrichment tests.

For the population frequency signature of differential selection, we use permutation tests to quantify the probability of having observed a mean heterozygosity per SNP hydrophobicity bin, described in detail above. The gene ontology and pathway enrichment tests use the g:SCS algorithm for correcting p-values for multiple-tests in the context of overlapping hierarchy of non-independent tests.

Computational packages used

All computations were done in Python 3.6 using the *numpy*, *scipy*, and *pandas* packages. All figures are made using the Python *matplotlib* package. The blobulation software is available at <https://github.com/BranniganLab/blobulator>.

Acknowledgments

The authors acknowledge the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey for providing access to the Amarel cluster and associated research computing resources that have contributed to the results reported here. The authors are also grateful to Ms. Kaitlin Bassi and Mr. Connor Pitman for useful discussions and comments on the manuscript, and to Dr. Cameron Abrams for coining the term “blobulation” to describe our algorithm.

Author Contributions

R.L. and G.B. established the blobulation approach. R.L. wrote the blobulation software and performed enrichment analyses. M.H. ran population-level analyses. All authors contributed to design of the study, discussion of results, generation of figures, and writing of the manuscript.

Competing Interests

The authors declare no competing financial interests.

References

- Cannon ME**, Mohlke KL. Deciphering the Emerging Complexities of Molecular Mechanisms at GWAS Loci. *Am J Hum Genet.* 2018 Nov; 103(5):637–653.
- 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 aug; 22(22):2729–2734. doi: 10.1093/bioinformatics/btl423.
- Capriotti E**, Fariselli P, Casadio R. A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics.* 2004 jul; 20(Suppl 1):i63–i68. doi: 10.1093/bioinformatics/bth928.

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

518 **Capriotti E**, Fariselli P. PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants.
519 *Nucleic Acids Research*. 2017 may; 45(W1):W247–W252. doi: 10.1093/nar/gkx369.

520 **Charlesworth B**, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular varia-
521 tion. *Genetics*. 1993 Aug; 134(4):1289–1303.

522 **Chen K**, Kurgan L, Ruan J. Optimization of the Sliding Window Size for Protein Structure Prediction. In: *2006*
523 *IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology* IEEE; 2006. doi:
524 [10.1109/cibcb.2006.330959](https://doi.org/10.1109/cibcb.2006.330959).

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

527 **Claussnitzer M**, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, Kathiresan S, Kenny EE, Lindgren CM,
528 MacArthur DG, North KN, Plon SE, Rehm HL, Risch N, Rotimi CN, Shendure J, Soranzo N, McCarthy MI. A brief
529 history of human disease genetics. *Nature*. 2020 jan; 577(7789):179–189. doi: 10.1038/s41586-019-1879-7.

530 **Conchillo-Solé O**, de Groot NS, Avilés FX, Vendrell J, Daura X, Ventura S. AGGRESKAN: a server for the predic-
531 tion and evaluation of "hot spots" of aggregation in polypeptides. *BMC bioinformatics*. 2007; 8(1):65. doi:
532 [10.1186/1471-2105-8-65](https://doi.org/10.1186/1471-2105-8-65).

533 **Cuanalo-Contreras K**, Mukherjee A, Soto C. Role of Protein Misfolding and Proteostasis Deficiency in
534 Protein Misfolding Diseases and Aging. *International Journal of Cell Biology*. 2013; 2013:1–10. doi:
535 [10.1155/2013/638083](https://doi.org/10.1155/2013/638083).

536 **Das RK**, Pappu RV. Conformations of intrinsically disordered proteins are influenced by linear sequence
537 distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences*. 2013 jul;
538 110(33):13392–13397. doi: [10.1073/pnas.1304749110](https://doi.org/10.1073/pnas.1304749110).

539 **Dickerson RE**. The structure of cytochrome c and the rates of molecular evolution. *J Mol Evol*. 1971 Mar;
540 1(1):26–45.

541 **DuBay KF**, Pawar AP, Chiti F, Zurdo J, Dobson CM, Vendruscolo M. Prediction of the Absolute Aggregation
542 Rates of Amyloidogenic Polypeptide Chains. *Journal of Molecular Biology*. 2004 aug; 341(5):1317–1326. doi:
543 [10.1016/j.jmb.2004.06.043](https://doi.org/10.1016/j.jmb.2004.06.043).

544 **Dyson HJ**, Wright PE. Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell*
545 *Biology*. 2005 mar; 6(3):197–208. doi: 10.1038/nrm1589.

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

548 **Gallagher MD**, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. *American Journal of Human*
549 *Genetics*. 2018; 102:717–730. doi: [10.1016/j.ajhg.2018.04.002](https://doi.org/10.1016/j.ajhg.2018.04.002).

550 **Hecht M**, Bromberg Y, Rost B. Better prediction of functional effects for sequence variants. *BMC Genomics*.
551 2015 jun; 16(S8). doi: 10.1186/1471-2164-16-s8-s1.

552 **Iqbal S**, Pérez-Palma E, Jespersen JB, May P, Hoksza D, Heyne HO, Ahmed SS, Rifat ZT, Rahman MS, Lage K,
553 Palotie A, Cottrell JR, Wagner FF, Daly MJ, Campbell AJ, Lal D. Comprehensive characterization of amino acid
554 positions in protein structures reveals molecular effect of missense variants. *Proceedings of the National*
555 *Academy of Sciences*. 2020 oct; p. 202002660. doi: [10.1073/pnas.2002660117](https://doi.org/10.1073/pnas.2002660117).

556 **Ittisoponpisan S**, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE. Can Predicted Protein 3D Struc-
557 tures Provide Reliable Insights into whether Missense Variants Are Disease Associated? *Journal of Molecular*
558 *Biology*. 2019 may; 431(11):2197–2212. doi: [10.1016/j.jmb.2019.04.009](https://doi.org/10.1016/j.jmb.2019.04.009).

559 **Jiang L**, Cao S, Cheung PPH, Zheng X, Leung CWT, Peng Q, Shuai Z, Tang BZ, Yao S, Huang X. Real-time monitoring
560 of hydrophobic aggregation reveals a critical role of cooperativity in hydrophobic effect. *Nature communi-*
561 *cations*. 2017 May; 8:15639. doi: 10.1038/ncomms15639.

562 **Karczewski KJ**, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birn-
563 baum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby
564 EG, Kosmicki JA, Walters RK, et al. Author Correction: The mutational constraint spectrum quantified from
565 variation in 141,456 humans. *Nature*. 2021 Feb; 590(7846):E53.

- 566 **Kyte J**, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular*
567 *Biology*. 1982 may; 157(1):105–132. doi: 10.1016/0022-2836(82)90515-0.
- 568 **Lins L**, Thomas A, Brasseur R. Analysis of accessible surface of residues in proteins. *Protein science : a publi-*
569 *cation of the Protein Society*. 2003 Jul; 12:1406–1417. doi: 10.1110/ps.0304803.
- 570 **Llaurens V**, Whibley A, Joron M. Genetic architecture and balancing selection: the life and death of differenti-
571 *ated variants*. *Mol Ecol*. 2017 May; 26(9):2430–2448.
- 572 **Lohia R**, Salari R, Brannigan G. Sequence specificity despite intrinsic disorder: How a disease-associated
573 *Val/Met polymorphism rearranges tertiary interactions in a long disordered protein*. *PLOS Computational*
574 *Biology*. 2019 oct; 15(10):e1007390. doi: 10.1371/journal.pcbi.1007390.
- 575 **López-Ferrando V**, Gazzo A, de la Cruz X, Orozco M, Gelpí JL. PMut: a web-based tool for the annotation of
576 *pathological variants on proteins, 2017 update*. *Nucleic Acids Research*. 2017 apr; 45(W1):W222–W228. doi:
577 10.1093/nar/gkx313.
- 578 **Ma Y**, Liu Y, Cheng J. Protein Secondary Structure Prediction Based on Data Partition and Semi-Random Sub-
579 *space Method*. *Scientific Reports*. 2018; 8(9856). doi: 10.1038/s41598-018-28084-8.
- 580 **Mir S**, Alhroub Y, Anyango S, Armstrong DR, Berrisford JM, Clark AR, Conroy MJ, Dana JM, Deshpande M, Gupta
581 *D*, Gutmanas A, Haslam P, Mak L, Mukhopadhyay A, Nadzirin N, Paysan-Lafosse T, Sehna D, Sen S, Smart OS,
582 *Varadi M, et al. PDBe: towards reusable data delivery infrastructure at protein data bank in Europe*. *Nucleic*
583 *Acids Research*. 2017 nov; 46(D1):D486–D492. doi: 10.1093/nar/gkx1070.
- 584 **Ng PC**, Henikoff S. Predicting Deleterious Amino Acid Substitutions. *Genome Research*. 2001 may; 11(5):863–
585 874. doi: 10.1101/gr.176601.
- 586 **Niroula A**, Urolagin S, Vihinen M. PON-P2: Prediction Method for Fast and Reliable Identification of Harmful
587 *Variants*. *PLOS ONE*. 2015 feb; 10(2):e0117380. doi: 10.1371/journal.pone.0117380.
- 588 **Panchenko AR**, Babu MM. Editorial overview: Linking protein sequence and structural changes to function
589 *in the era of next-generation sequencing*. *Current Opinion in Structural Biology*. 2015 jun; 32:viii–x. doi:
590 10.1016/j.sbi.2015.06.005.
- 591 **Park Y**, Hayat S, Helms V. Prediction of the burial status of transmembrane residues of helical membrane
592 *proteins*. *BMC Bioinformatics*. 2007; 8(1):302. doi: 10.1186/1471-2105-8-302.
- 593 **Parthiban V**, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations. *Nu-*
594 *cleic Acids Research*. 2006 jul; 34(Web Server):W239–W242. doi: 10.1093/nar/gkl190.
- 595 **Patel A**, Lee HO, Jawerth L, Maharana S, Jahnel M, Hein MY, Stoyanov S, Mahamid J, Saha S, Franzmann TM,
596 *Pozniakovski A, Poser I, Maghelli N, Royer LA, Weigert M, Myers EW, Grill S, Drechsel D, Hyman AA, Alberti S.*
597 *A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation*. *Cell*. 2015 aug;
598 162(5):1066–1077. doi: 10.1016/j.cell.2015.07.047.
- 599 **Pierron D**, Cortés NG, Letellier T, Grossman LI. Current relaxation of selection on the human genome: tolerance
600 *of deleterious mutations on olfactory receptors*. *Mol Phylogenet Evol*. 2013 Feb; 66(2):558–564.
- 601 **Popov P**, Bizin I, Gromiha M, A K, Frishman D. Prediction of disease-associated mutations in the transmem-
602 *brane regions of proteins with known 3D structure*. *PLOS ONE*. 2019 jul; 14(7):e0219452. doi: 10.1371/jour-
- 603 *nal.pone.0219452*.
- 604 **Prlić A**, Kalro T, Bhattacharya R, Christie C, Burley SK, Rose PW. Integrating genomic information with protein se-
605 *quence and 3D atomic level structure at the RCSB protein data bank*. *Bioinformatics*. 2016 aug; 32(24):3833–
606 3835. doi: 10.1093/bioinformatics/btw547.
- 607 **Raudvere U**, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J. g: Profiler: a web server for functional
608 *enrichment analysis and conversions of gene lists (2019 update)*. *Nucleic acids research*. 2019; 47(W1):W191–
609 W198.
- 610 **Richards S**, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Others.
611 *Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of*
612 *the American College of Medical Genetics and Genomics and the Association for Molecular Pathology*. *Genet*
613 *Med*. 2015; 17(5):405–423.

- 614 **Rose PW**, Prlić A, Altunkaya A, Bi C, Bradley AR, Christie CH, Costanzo LD, Duarte JM, Dutta S, Feng Z, et al.
615 The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic acids*
616 *research*. 2016 jan; 45(D1):D271–D281.
- 617 **Sander O**, Sommer I, Lengauer T. Local protein structure prediction using discriminative models. *BMC Bioin-*
618 *formatics*. 2006; 7(1):14. doi: 10.1186/1471-2105-7-14.
- 619 **Schlessinger A**, Rost B. Protein flexibility and rigidity predicted from sequence. *Proteins: Structure, Function,*
620 *and Bioinformatics*. 2005 aug; 61(1):115–126. doi: 10.1002/prot.20587.
- 621 **Siepel A**, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards
622 S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. Evolutionarily conserved elements in
623 vertebrate, insect, worm, and yeast genomes. *Genome Res*. 2005 Aug; 15(8):1034–1050.
- 624 **Somel M**, Wilson Sayres MA, Jordan G, Huerta-Sanchez E, Fumagalli M, Ferrer-Admetlla A, Nielsen R. A scan
625 for human-specific relaxation of negative selection reveals unexpected polymorphism in proteasome genes.
626 *Mol Biol Evol*. 2013 Aug; 30(8):1808–1815.
- 627 **Stone EA**. Physicochemical constraint violation by missense substitutions mediates impairment of protein
628 function and disease severity. *Genome Research*. 2005 jun; 15(7):978–986. doi: 10.1101/gr.3804205.
- 629 **Sunyaev SR**, Lathe WC 3rd, Ramensky VE, Bork P. SNP frequencies in human genes an excess of rare alleles
630 and differing modes of selection. *Trends Genet*. 2000 Aug; 16(8):335–337.
- 631 **Tartaglia GG**, Vendruscolo M. The Zyggregator method for predicting protein aggregation propensities. *Chem-*
632 *ical Society Reviews*. 2008; 37:1395–1401. doi: 10.1039/b706784b.
- 633 **Teng S**, Srivastava AK, Wang L. Sequence feature-based prediction of protein stability changes upon amino
634 acid substitutions. *BMC Genomics*. 2010; 11(Suppl 2):S5. doi: 10.1186/1471-2164-11-s2-s5.
- 635 **Thomas PD**, Kejariwal A. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian
636 disease: Evolutionary evidence for differences in molecular effects. *Proceedings of the National Academy of*
637 *Sciences*. 2004 oct; 101(43):15398–15403. doi: 10.1073/pnas.0404380101.
- 638 **Tovo-Rodrigues L**, Recamonde-Mendoza M, Paixão-Côrtés VR, Bruxel EM, Schuch JB, Friedrich DC, Rohde LA,
639 Hutz MH. The role of protein intrinsic disorder in major psychiatric disorders. *American Journal of Medical*
640 *Genetics Part B: Neuropsychiatric Genetics*. 2016 may; 171(6):848–860. doi: 10.1002/ajmg.b.32455.
- 641 **Uversky VN**. Unusual biophysics of intrinsically disordered proteins. *Biochimica et Biophysica Acta (BBA) -*
642 *Proteins and Proteomics*. 2013 may; 1834(5):932–951. doi: 10.1016/j.bbapap.2012.12.008.
- 643 **Uversky VN**. Intrinsically disordered proteins and their (disordered) proteomes in neurodegenerative disor-
644 ders. *Frontiers in Aging Neuroscience*. 2015 mar; 7. doi: 10.3389/fnagi.2015.00018.
- 645 **Uversky VN**. Intrinsically Disordered Proteins and Their “Mysterious” (Meta)Physics. *Frontiers in Physics*. 2019
646 feb; 7. doi: 10.3389/fphy.2019.00010.
- 647 **Uversky VN**, Iakoucheva LM, Dunker AK. Protein Disorder and Human Genetic Disease. In: eLS John Wiley and
648 Sons, Ltd: Chichester. 2012 feb; doi: 10.1002/9780470015902.a0023589.
- 649 **Visscher PM**, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 Years of GWAS Discovery: Biology,
650 Function, and Translation. *Am J Hum Genet*. 2017 Jul; 101(1):5–22.
- 651 **Wainreb G**, Ashkenazy H, Bromberg Y, Starovolsky-Shitrit A, Haliloglu T, Ruppini E, Avraham KB, Rost B, Ben-Tal
652 N. MuD: an interactive web server for the prediction of non-neutral substitutions using protein structural
653 data. *Nucleic Acids Research*. 2010 jun; 38(suppl_2):W523–W528. doi: 10.1093/nar/gkq528.
- 654 **Wang Y**, Mao H, Yi Z. Protein secondary structure prediction by using deep learning method. *Knowledge-Based*
655 *Systems*. 2017; 118:115–123. doi: 10.1016/j.knsys.2016.11.015.
- 656 **Ward JJ**, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and Functional Analysis of Native Disorder in
657 Proteins from the Three Kingdoms of Life. *Journal of Molecular Biology*. 2004 mar; 337(3):635–645. doi:
658 10.1016/j.jmb.2004.02.002.
- 659 **Weinreb PH**, Zhen W, Poon AW, Conway KA, Lansbury PT. NACP, A Protein Implicated in Alzheimer's Disease
660 and Learning, Is Natively Unfolded†. *Biochemistry*. 1996 jan; 35(43):13709–13715. doi: 10.1021/bi961799n.

- 661 **Yang Y**, Gao J, Wang J, Heffernan R, Hanson J, Paliwal K, Zhou Y. Sixty-five years of the long march in protein
662 secondary structure prediction: the final stretch? *Briefings in Bioinformatics*. 2018 may; 19(3):482–494. doi:
663 10.1093/bib/bbw129.
- 664 **Yip YL**, Famiglietti M, Gos A, Duek PD, David FPA, Gateau A, Bairoch A. Annotating single amino acid
665 polymorphisms in the UniProt/Swiss-Prot knowledgebase. *Human Mutation*. 2008; 29(3):361–366. doi:
666 [10.1002/humu.20671](https://doi.org/10.1002/humu.20671).
- 667 **Zhang B**, Li J, Lü Q. Prediction of 8-state protein secondary structures by a novel deep learning architecture.
668 *BMC Bioinformatics*. 2018 aug; 19(1). doi: 10.1186/s12859-018-2280-5.