PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Contiguously hydrophobic sequences are functionally significant throughout the human exome

Ruchi Lohia[a,b], Matthew E. B. Hansen[c,1], and Grace Brannigan[a,d,1,2]

Edited by Ken Dill, Stony Brook University, Stony Brook, NY; received September 12, 2021; accepted February 2, 2022

Hydrophobic interactions have long been established as essential for stabilizing structured proteins as well as drivers of aggregation, but the impact of hydrophobicity on the functional significance of sequence variants has rarely been considered in a genome-wide context. Here we test the role of hydrophobicity on functional impact across 70,000 disease- and non–disease-associated single-nucleotide polymorphisms (SNPs), using enrichment of disease association as an indicator of functionality. We find that functional impact is uncorrelated with hydrophobicity of the SNP itself and only weakly correlated with the average local hydrophobicity, but is strongly correlated with both the size and minimum hydrophobicity of the contiguously hydrophobic sequence (or "blob") that contains the SNP. Disease association is found to vary by more than sixfold as a function of contiguous hydrophobicity parameters, suggesting utility as a prior for identifying causal variation. We further find signatures of differential selective constraint on hydrophobic blobs and that SNPs splitting a long hydrophobic blob or joining two short hydrophobic blobs are particularly likely to be disease associated. Trends are preserved for both aggregating and nonaggregating proteins, indicating that the role of contiguous hydrophobicity extends well beyond aggregation risk.

sequence–function relationship | protein hydrophobicity | single-nucleotide polymorphism | population genetics | computational methods

Significance

Proteins rely on the hydrophobic effect to maintain structure and interactions with the environment. Surprisingly, natural selection on amino acid hydrophobicity has not been detected using modern genetic data. Analyses that treat each amino acid separately do not reveal significant results, which we confirm here. However, because the hydrophobic effect becomes more powerful as more hydrophobic molecules are introduced, we tested whether unbroken stretches of hydrophobic amino acids are under selection. Using genetic variant data from across the human genome, we find evidence that selection increases with the length of the unbroken hydrophobic sequence. These results could lead to improvements in a wide range of genomic tools as well as insights into protein-aggregation disease etiology and protein evolutionary history.

Protein structure is commonly understood to mediate the effects of sequence on function, but 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 (1–5) and sensitive to sequence (6–11). A missing framework for sequence modularity and organization presents a conceptual and technical barrier to understanding the underlying mechanisms of sequence dependence, as well as the genetic basis for heritable traits and disease risks. Structured proteins are clearly modular, but identifying the module boundaries has required explicit knowledge or prediction of secondary structure. There has been no generic approach for detecting organization when structure is unknown and unpredictable or nonexistent. Here, we propose a general role for contiguous stretches of hydrophobic residues in organizing sequences and determining sequence sensitivity.

We showed in a previous study that a long intrinsically disordered protein can retain modularity of tertiary interactions, despite the absence of tertiary structure. Fully atomistic, explicit-solvent molecular dynamics simulations of the 91-residue disordered prodomain of Brain Derived Neurotrophic Factor (BDNF) revealed a soft network of tertiary contacts between contiguous stretches of hydrophobic residues (12). To distinguish these stretches of hydrophobic residues from any other more traditional segment or domain definition, we follow terminology common to polymer physics and call these stretches "blobs." Blobs may contain secondary structure elements, but are not required to do so. These results suggested a more generic framework for interaction-based functional modularity, which could be determined directly from sequence. However, the usefulness of this approach had not been tested beyond the single protein in which it was developed.

Many variant-to-function prediction methods rely on some form of residue characterization of the variant and its local sequence. In addition to physicochemical properties (13–17) these may include evolutionary conservation (13, 14, 16, 18–21) and structural propensities (17, 22–27). Such methods may also rely on known protein structures to incorporate properties such as local secondary structure and solvent accessibility (27). Fewer than 35% of human protein-coding genes have structures deposited in the protein data bank (28), and complete structures have been experimentally solved for a tiny fraction of known proteins (29, 30).

In the absence of structural information, physicochemical properties like hydrophobicity can still be determined from sequence, but the properties of individual residues are not

Author contributions: R.L., M.E.B.H., and G.B. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]M.E.B.H. and G.B. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: grace.brannigan@rutgers.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2116267119/-/DCSupplemental.

Published March 16, 2022.

predictive. Most attempts at incorporating the local sequence have used a fixed-width sliding window centered around the SNP, which neglects the intrinsic modularity of protein sequences and may contribute to the relative weakness of these approaches. For example, if the mutated residue is near the module boundary, the mutation-centered window will partially overlap a module that does not contain the SNP. We propose here that the blob surrounding a residue provides a natural and sequence-informed definition for the local sequence context. Furthermore, due to the cooperativity of the hydrophobic effect (31) and the tendency for hydrophobic residues to be buried (32), we hypothesized that hydrophobic blobs ("h-blobs") would form interaction-rich and mutation-sensitive clusters across a generic proteome.

In the present work, we detect h-blobs across the human proteome and characterize their structural properties and functional impact. We calculate the distribution of secondary structures for residues inside and outside h-blobs and use solvent-accessibility calculations to test our hypothesis that h-blobs represent buried regions of solvated proteins. Our analysis of function uses disease association as a proxy for functional impact. We test for enrichment of disease-associated SNPs as a function of hydrophobic blob length, residue hydrophobicity, and average hydrophobicity of a window centered around the SNP. We complement these results using population frequencies to test for signatures of selective constraint in h-blobs. Furthermore, we consider functional impact for multiple special cases, including transmembrane regions, aggregating proteins, and SNPs that split or merge blobs or change specific blob properties. In sum, we present several signatures consistent with the hypothesis that proteins are partly organized around domains of contiguous hydrophobicity and demonstrate that we can roughly delineate such regions from sequence alone.

Results

Regions of Contiguous Hydrophobicity Are Enriched for β-Strands and Buried Residues. The blobulation algorithm, first presented in ref. 12, is a method for tunable segmentation or edge detection in protein sequences based on hydrophobicity. The original "whole-sequence" blobulation algorithm digitizes the sequence and then clusters it. In the preliminary digitization step (Fig. 1*A*), each residue is classified as hydrophobic or nonhydrophobic, depending on whether the Kyte–Dolittle hydropathy score (33) is above or below the hydrophobicity threshold $H^\star$. In the main clustering step (Fig. 1*B*), the algorithm scans the given amino acid sequence to identify stretches of at least $L_{\min}$ sequential residues that were classified as hydrophobic during digitization. These stretches are called hydrophobic blobs or h-blobs.

The residues that are not assigned to h-blobs will contain a mix of nonhydrophobic residues and isolated hydrophobic residues, in stretches that either link h-blobs or terminate the sequence: Short stretches that contain fewer than $L_{\min}$ residues are classified as separator blobs ("s-blobs"), while long stretches with at least $L_{\min}$ residues are classified as polar blobs ("p-blobs"). The blobs can then be characterized based on any collective property of the blob sequence; the h, p, and s designations also constitute the primary blob characterization termed the "hydrophobicity class." In ref. 12 we also introduced a higher-order classification: Adjacent h-blobs that were separated only by the short s-blobs were called "h-groups." We include this designation in Fig. 1*B* for completeness but do not explicitly consider h-groups in this paper. However, h-groups are captured implicitly by variation of the $H^\star$ threshold; the long h-blobs that are detected at low $H^\star$ would be classified as h-groups at high $H^\star$.

The blobulation algorithm can be used in multiple ways, and we apply it in three slightly different ways within this study. In the original "whole sequence" version just described, the entire sequence is unambiguously and completely decomposed into h-, p-, and s-blobs using two fixed parameters $\{H^\star, L_{\min}\}$ provided by the user. In this section we continue to use this approach, because the analysis considers whole proteins. Alternatively, we can fix one parameter and determine the value of the second parameter that would assign a fixed residue to an h-blob. In subsequent sections we use two such variations ("unconstrained length" and "unconstrained threshold") for analyzing the blob properties surrounding specific sets of SNPs. Both approaches are described further at first use.

We find that nearly half of the residues in the proteome meet our original (12) relaxed criteria for h-blobs, which include short, moderately hydrophobic sequences. More specifically, using $L_{\min} = 4$ and $H^\star = 0.4$, the residues in the Universal Protein Resource (UniProt) database (36) (n = 6,459 proteins) are distributed as follows: 45% in h-blobs, 52% in p-blobs, and 3% in s-blobs. Stricter criteria can be used to isolate the long, highly hydrophobic blobs that cover less than 10% of the proteome: Using $H^\star = 0.5$ and $L_{\min} = 8$, the distribution is 7% in h-blobs, 93% in p-blobs, and <1 % in s-blobs. The effect of varying these two parameters is shown for cytochrome C peroxidase in Fig. 1*C*; as the criteria are made more restrictive, the algorithm isolates one long and very hydrophobic blob at the core of the protein.

To test the hypothesis that h-blobs would be buried in globular, structured proteins, we calculated the relative solvent-accessibility surface area (SASA) for each blob type, determined using structures in the Protein Data Bank (PDB) (excluding transmembrane domains). Relaxed criteria ($L_{\min} = 4$, $H^\star = 0.4$) were used for this calculation to maximize the amount of available data and to keep the analysis conservative (Dataset S1). These calculations (Fig. 1*D*) confirmed that residues in h-blobs have a substantially lower SASA (0.21) than p-blobs (0.35) or s-blobs (0.33); SE for all three quantities is less than 0.001. Together, these results suggest that h-blobs condense into buried clusters that are rich in intra- or interprotein interactions. As illustrated in Fig. 1*C*, we expect this difference to be even larger for h-blobs that meet stricter criteria.

The overlap between blob hydrophobicity class and secondary structure was measured by blobulating all unique proteins in the PDB and tabulating the fraction of residues occurring in helices (α, 3–10, π), strand (β-bridge or extended β-strand), or coil for each blob type (Dataset S2). As shown in Fig. 1*E*, all blob types contain a comparable fraction of helices, although the fraction in h-blobs is slightly greater than in non–h-blobs. We do not expect contiguous hydrophobicity to be particularly correlated with helical structure (with the exception of transmembrane helices): Helices frequently have multiple faces with differential solvent accessibility, and the sequence needs to cycle through residues that are appropriate for each face. H-blobs, however, are about twice as likely to contain strands as non–h-blobs (Fig. 1*E*). β-strands are secondary structure elements, but they are also indicators of tertiary interactions, since β-strands will have a pairing β-strand. These results are consistent with our first application of blobulation to simulated conformations of the long disordered pro region of BDNF (12), which revealed a network of soft tertiary interactions mediated through pairing of β-strands in h-blobs.

As an example, Fig. 1*E* also shows the blob assignments for the ubiquitin sequence mapped onto its structure. The N-terminal β-hairpin is assigned to two h-blobs, separated by an s-blob that is adjacent to the turn. The C-terminal β-strand is also assigned to an h-blob, capped by an s-blob at the terminus. Finally, the

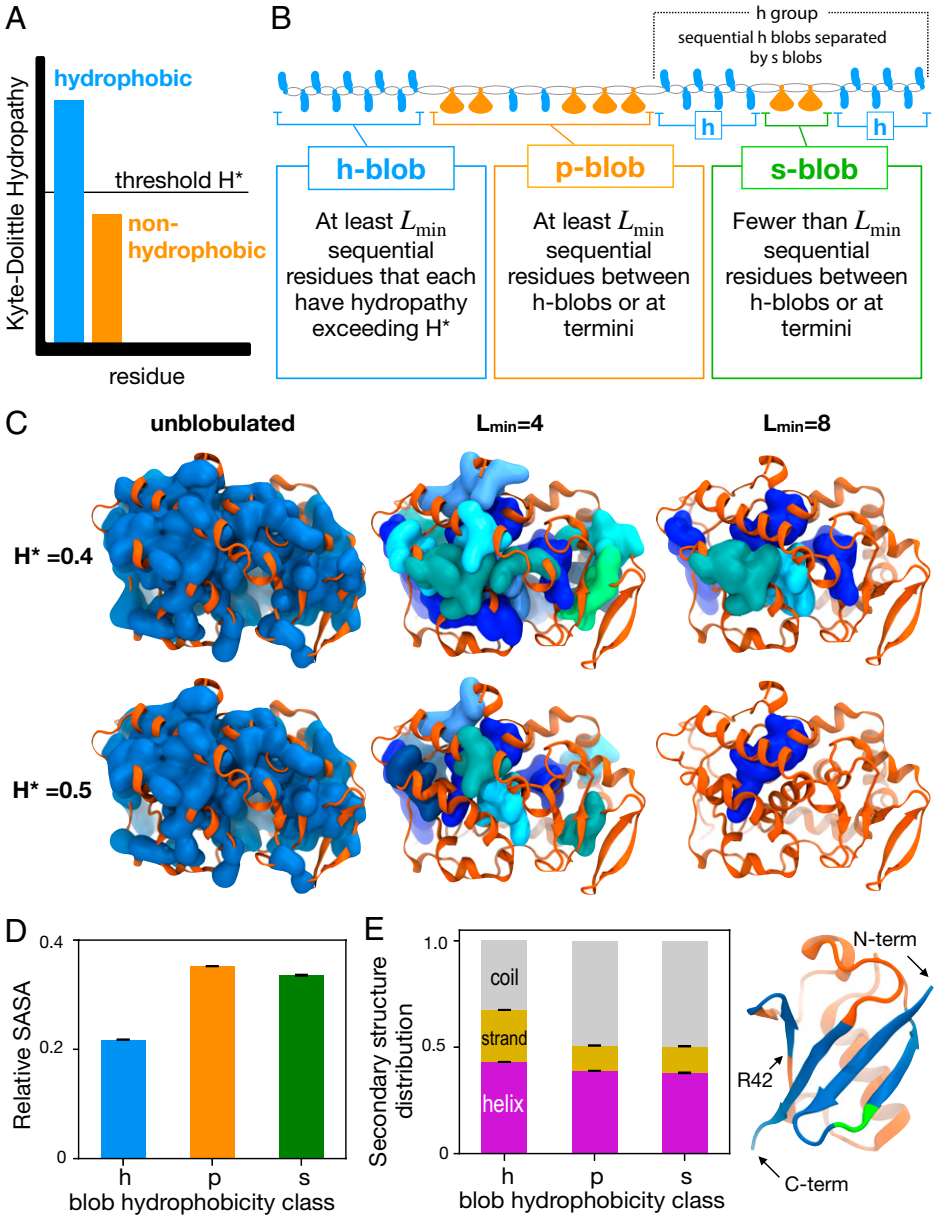

Fig. 1. Whole-sequence blobulation algorithm for segmentation of proteins. (*A*) First, the sequence is digitized: Residues are classified as hydrophobic or nonhydrophobic depending on whether they have a Kyte–Dolittle (33) hydropathy falling above or below the user-provided threshold H^*, respectively. (*B*) The clustering step acts on the digitized sequence, which is illustrated here as a cartoon: Residues above and below the H^* threshold are shown as blue ovals and orange fans, respectively. The clustering step scans the digitized sequence according to the indicated criteria, first detecting h-blobs, then p-blobs, and finally s-blobs. L_{min} is the user-provided minimum blob size. The blobulation outcome for this particular chain would be valid for $2 < L_{min} < 6$. The software and a web interface are freely available as described in *Materials and Methods*. (*C*) Cytochrome C peroxidase (2CYP) (34) shown for two different digitization thresholds H^* (rows) and three different clustering criteria (columns), including no clustering (*Left* "unblobulated" column) or two different values of L_{min} (*Center* and *Right* columns, respectively). The blue surface of the unblobulated sequence depicts canonical hydrophobic residues; the $H^* = 0.4$ row also includes serine and threonine, which have Kyte–Dolittle hydropathy scores of 0.41 and 0.42, respectively. H-blobs in *Center* and *Right* columns are colored by arbitrarily varying shades of blue to distinguish individual blobs. (*D*) Relative SASA for nontransmembrane blobs in the PDB, categorized by blob hydrophobicity class. (*E*) *Left*, distribution of secondary structures for sequences with structures in the PDB, categorized by blob hydrophobicity class. *Right*, blobulated ubiquitin (PDB ID 5GO7) (35), colored by blob hydrophobicity class, as in *B*. In both *D* and *E*, the blobulation algorithm uses $L_{min} = 4$ and $H^* = 0.4$; error bars are SEs ($n > 500,000$ for h- and p-blobs, $n > 50,000$ for s-blobs). See *Materials and Methods* for details.

Pro37-Ala46 β-strand is a p-blob from Pro37 to Arg42 and then switches to an h-blob as it crosses into the h-blob–rich part of the protein. That h-blob continues until the chain bends back toward the other p-blobs. This example suggests that while blob boundaries can align with secondary structure elements, they are more fundamentally correlated with the location of the blob within the three-dimensional protein structure.

Average Local Hydrophobicity Is a Weak Indicator of Disease Association. To test whether the residue hydrophobicity is prima facie correlated with functional impact, we calculated the enrichment of disease-associated missense SNPs ("dSNPs") as a function of hydrophobicity of the reference allele. Throughout this paper, unless otherwise noted, dSNPs are tested for enrichment relative to the expectation set by missense SNPs that are not disease-associated ("nSNPs"). For example, the phrase "dSNPs are enriched in blobs of type X" means that the proportion of dSNPs found in blobs of type X is larger than the proportion of nSNPs found in blobs of type X. As shown in Fig. 2*B*, we did not detect any significant correlation (Pearson's r = 0.02,

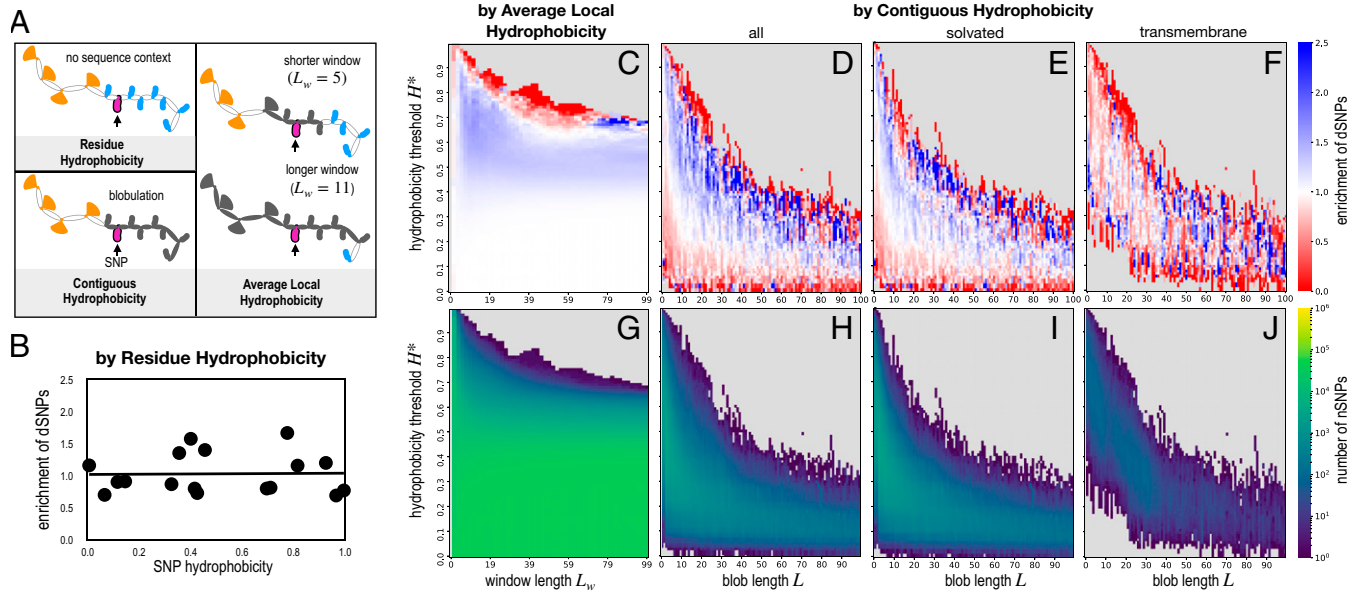

Fig. 2. Effect of segmentation approach, length, hydrophobicity threshold, and solvation on calculated enrichment of dSNPs in hydrophobic segments. (*A*) Illustration of three measures of SNP hydrophobicity (residue, contiguous, and average local) for the indicated SNP, found within a hypothetical peptide chain composed of residues classified as hydrophobic (blue ovals) or nonhydrophobic (orange fans) for a given $H^\star$, as in Fig. 1*B*. Unconstrained-length blobulation determines the local sequence (shaded in gray) by detecting contiguous hydrophobic residues, which together form an h-blob of length L; the moving-window approach determines the local sequence using a fixed number L_w of residues centered around the SNP. (*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*–*F*) Enrichment of dSNPs in hydrophobic segments, as a function of segment length and threshold, for (*C*) fixed-length hydrophobic windows of length L_w in which the average hydrophobicity is above $H^\star$, and (*D*–*F*) h-blobs of length L, calculated with the threshold $H^\star$ for (*D*) all SNPs, (*E*) those outside of transmembrane domains, and (*F*) those in transmembrane domains. (*G*–*J*) The total number of nSNPs per bin for the corresponding enrichment heatmaps in *C*–*F* (e.g., *I* shows the nSNP counts for the enrichments in *E*). Each panel (*C*–*J*) is colored according to the scale at *Right* end of the row, and bins with no data are colored gray.

$n = 17$) between lone SNP hydrophobicity and dSNP enrichment, meaning that the hydrophobicity of a residue considered in isolation does not show this particular signature of functionality.

The effect of average local hydrophobicity on the enrichment of dSNPs was calculated using moving windows of length L_w centered around each SNP. While there is no "standard" window size, most SNP prediction programs use a window size in the range of 1 to 21 residues (16, 20, 37, 38). The window size is chosen to balance concerns that small window sizes may not accurately capture the "local" sequence (39–41) whereas large window sizes can decrease the signal-to-noise ratio (42). Here we computed the mean window hydrophobicity $\bar{H}_i$ for all SNPs i in our SNP dataset, while also varying L_w. Fig. 2*C* shows the enrichment of dSNPs for which $\bar{H}_i > H^\star$, for the range of moving-window widths $L_\mathrm{w} = 1$ to 99. As is evident in Fig. 2*C*, the enrichment of dSNPs is relatively insensitive to the window size for the regime where $L_\mathrm{w} \geq 6$ and $H^\star \leq 0.65$. The total count of nSNPs in each bin is shown in Fig. 2*G* and was similarly insensitive to window size for larger thresholds. These results suggest that distant residues introduce noise that averages out in a proteome-wide analysis, but their inclusion in the window would still reduce precision for any individual SNP.

Surprisingly, we detect a narrow band of dSNP depletion for windows with a high average hydrophobicity (the red signal for $H^\star \geq 0.65$ in Fig. 2*C*). This signal is due to only a handful of SNPs (see the counts in Fig. 2*G*), but we observe a similar pattern using the blobulation algorithm, and we discuss its origins in the next sections.

Contiguous Hydrophobicity Is a Strong Indicator of Disease Association. The use of a fixed-width window neglects the inherent dispersion in the size of protein modules, which are captured using blobulation (Fig. 2*A*). To quantify the blob properties for each SNP, we used a blobulation variant we call unconstrained-length blobulation, which fixes the threshold $H^\star$ and a reference residue i but imposes no minimum blob length. This approach first tests whether the hydropathy score for residue i exceeds $H^\star$, and if so, it calculates the exact length L of the h-blob that contains residue i. Unconstrained-length blobulation is formally equivalent to whole-sequence blobulation with $L_\mathrm{min} = 1$, but is more efficient since we are analyzing only the relevant part of the sequence.

Specifically, we applied unconstrained-length blobulation to each SNP in the dataset (using the reference allele) and a given hydrophobicity threshold $H^\star$ and then determined L. We repeated this calculation for a series of $H^\star$ values, and for dSNPs and nSNPs separately (Dataset S3), and then tabulated the proportion of dSNPs and nSNPs in each $(H^\star, L)$ bin (Dataset S4). The resulting enrichment of dSNPs as a function of blob length L and threshold $H^\star$ is shown in Fig. 2*D*, and the total number of nSNPs per bin is shown in Fig. 2*H*. We observe a consistent relationship between hydrophobicity of the local blob and dSNP enrichment. dSNPs are depleted in weakly hydrophobic blobs, are neutral for moderately hydrophobic blobs, and become more enriched as the blob gets longer and/or satisfies a stricter hydrophobicity threshold. The trend is primarily monotonic, which supports the hypothesis that hydrophobic blobs constitute functional elements.

We do find an exception to the trend at the plot boundary: Blobs that satisfied the very strictest criteria were depleted in dSNPs, consistent with the results using the moving-window analysis. The depletion signal persisted even when bins with very few samples were removed. In the next section, we consider two potential reasons for this depletion: 1) dSNPs in these blobs are so deleterious that they are selected out of the population or 2) some subset of nSNPs is functional and under balancing selection or relaxed constraint. In addition to a consistent trend, the analysis returns a large spread in enrichment/depletion values: 3.2% of the bins in Fig. 2*D* are significantly depleted below 0.5, while

13% have a significant enrichment of greater than 1.5, and 3.5% have a significant enrichment greater than 3 (significance based on binomial test, $P < 10^{-3}$; *Materials and Methods*). This range indicates that hydrophobicity-based sequence segmentation could be particularly useful for assessing the riskiness of SNPs located in long or very hydrophobic sequences. The numerical enrichment values for each bin are provided in Dataset S4.

Transmembrane helices will intrinsically require contiguous residues that are at least moderately hydrophobic (with the exception of pore-lining helices) and it seemed possible that our results were dominated by the distinctive properties of such transmembrane domains. We further decomposed the data into contributions from SNPs that are not in transmembrane domains (Fig. 2*E*) and those that are (Fig. 2*F*). The counts of nSNPs per bin for each case are shown in Fig. 2 *I* and *J*, respectively, indicating that the overall dataset includes relatively few SNPs in transmembrane domains. We find that enrichment trends for solvated dSNPs (Fig. 2*E*) mimic trends in the combined dataset (Fig. 2*D*). Thus, we conclude that enrichment of dSNPs in h-blobs is a general trend rather than an indication of membrane exposure.

We note that distinguishing between solvated and transmembrane SNPs also suggests appropriate blobulation parameters for transmembrane domains. Transmembrane helices are known to be 24 residues long on average, with about 19 residues forming the hydrophobic core (43). As expected, none of the transmembrane residues are found in h-blobs that are much shorter than 19 residues (Fig. 2*J*), except when the threshold is sufficiently high to exclude those polar residues that are frequently found in transmembrane helices. More specifically, raising the threshold beyond $H^\star = 0.36$ typically excludes all charged and polar residues but serine and threonine; further raising the threshold beyond $H^\star = 0.42$ typically excludes all charged and polar residues. Any transmembrane polar residues that fall below the $H^\star$ threshold will divide the transmembrane into multiple h-blobs, which is consistent with the large number of short h-blobs at high thresholds. In contrast, when short polar and charged linkers between transmembrane segments are included ($H^\star < 0.3$), multiple transmembrane segments may be grouped into the same blob, with a minimum length around 20 residues. Such a series of transmembrane segments would also be a common example of the h-group illustrated in Fig. 1*B*, although we do not use that hierarchical descriptor in this analysis.

The Genetic Diversity of Disease-Associated Variants Is Lowest in the Most Hydrophobic Blobs. If hydrophobic blobs capture functional segments of coding regions, then we may expect to see signatures of differential selective constraint with varying blob hydrophobicity. We test this hypothesis by examining whether or not the genetic diversity of a SNP varies with the surrounding blob hydrophobicity. Blobulation approaches with a fixed threshold H^* will not distinguish between blobs that barely exceed the threshold and those that significantly exceed it, so here we use unconstrained-threshold blobulation. In this blobulation variant, we fix the minimum length $L_{\min}$ but do not fix the hydropathy threshold $H^\star$. Instead, for a given residue i, we calculate $H_{\max}$: the maximum possible value of $H^\star$ that would still assign residue i to an h-blob that is at least $L_{\min}$ residues long. Here we use $L_{\min} = 4$.

A benchmark measure of genetic diversity is the expected heterozygosity $\Theta \equiv 2\nu(1-\nu)$, where ν is the frequency of the coded allele. Sequences under more functional constraint, like exons in essential proteins, experience purifying selection (removal of almost all new functional alleles), which lowers Θ compared to regions under little or no constraint (44–48). Conversely, balancing selection (maintenance of multiple functional alleles) causes increased genetic diversity over a genomic region (49). Population substructure can also increase the genetic diversity, but such effects would occur genome-wide and would not be correlated with blob hydrophobicity class.

The results of stratifying the genetic diversity of SNPs by their surrounding maximum blob hydrophobicity are shown in Fig. 3. The population frequencies are based on non-Finnish Europeans (50) (Dataset S5), for which the UniProt dSNP and nSNP functional annotations are expected to be reasonably accurate. The average heterozygosity of nSNPs in low-hydrophbicity domains ($H_{\max} \leq 0.25$) is 0.125 (Fig. 3*A*), while that for dSNPs is 40 times lower, 0.003 (Fig. 3*B*). For both nSNPs and dSNPs, the genetic diversity shows a significant departure from the neutral expectation (outside the 1st to 99th null percentiles) for the largest $H_{\max}$ bin, $H_{\max} \geq 0.75$. For nSNPs, the heterozygosity is larger than expected (>99th null percentile), indicative of balancing selection. In contrast, for dSNPs the heterozygosity is smaller than expected (<1st null percentile), indicative of greater purifying selection in highly hydrophobic blobs. In other words, the disease-associated alleles in highly hydrophobic blobs appear to be more deleterious than disease-associated alleles in lower-hydrophobicity blobs.

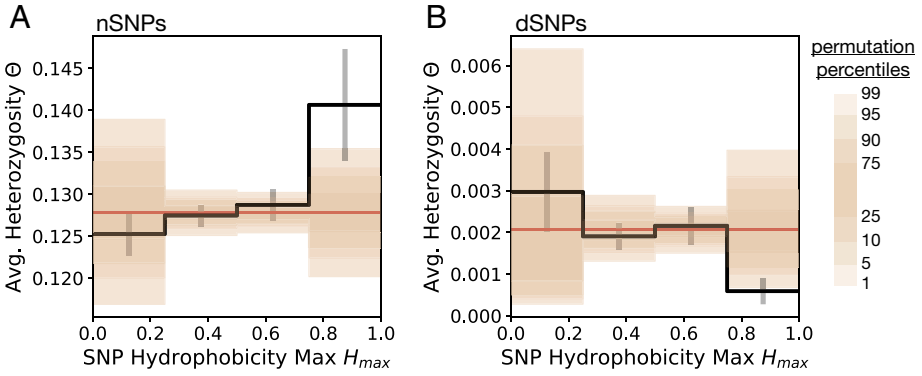

Fig. 3. Expected heterozygosity of SNPs in Europeans as a function of blob hydrophobicity. (*A* and *B*) The black line shows the average expected heterozygosity Θ for (*A*) nSNPs and (*B*) dSNPs in h-blobs, determined using unconstrained-threshold blobulation ($L_{\min} = 4$) and binned by the resulting maximum SNP hydrophobicity $H_{\max}$. $H_{\max}$ bins have width $\Delta H_{\max} = 0.25$ and error bars represent SEs in the mean for that bin. Only SNPs where at least one of two alleles are in an annotated h-blob are considered. Frequencies are from the gnomAD cohort of non-Finnish Europeans (50). The horizontal brown line is the average heterozygosity for the SNPs in each panel. The background null distribution is generated from random permutation of frequency assignment among nSNPs and among dSNPs, respectively, and is shown as brown shaded regions, with the null percentiles shown in the key at *Right*.

The results here may explain the nonmonotonic behavior of the dSNP-to-nSNP enrichment seen in Fig. 2 *C* and *D* for both the moving-window and blobulation 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 whether the lack of dSNPs relative to nSNPs is due to increased selection against dSNPs, increased balancing selection for nSNPs, 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 Fig. 2 *C* and *D* in the high-hydrophobicity regions is caused by both increased selection against disease-associated alleles and increased selection for the maintenance of polymorphisms in the non–disease-associated variants.

We tested whether the trends observed in Fig. 3 are specific to non-Finnish Europeans by performing the same analysis based on East Asian population frequencies (*SI Appendix*, Fig. S1 and Dataset S5). The East Asian frequency dataset contains an order of magnitude fewer individuals than the non-Finnish European data, $n \sim 1,500$ vs. $n \sim 32,000$ individuals, which means the frequency resolution is worse than for the non-Finnish European data. Consequently, we do not observe the same level of statistical significance, but 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 $H_{\max}$. The stratification of genetic diversity with blob properties appears to be a feature shared across diverse human populations.

For additional context on the proteins containing the blobs in the $H_{\max} \geq 0.75$ bin, we use gene-ontology and pathway enrichment tests (*Materials and Methods* and Dataset S6). 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 $N = 1,803$ 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, 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 signal transduction. This accords with previous observations that olfactory-related genes exhibit signs of balancing selection (51) and/or relaxed purifying selection (52). This analysis has essentially reidentified this same feature of greater diversity in chemical sensing pathways, found here because the increased genetic diversity is specifically present in highly hydrophobic subdomains of those proteins. In contrast, the $H_{\max} \geq 0.75$ dSNPs are enriched for cation and small molecule transmembrane transport proteins. These SNPs reside in proteins enriched for critical membrane-bound transporters, consistent with the signal of higher functional constraint.

dSNP Enrichment Emerges in Shorter and More Weakly Hydrophobic Blobs for Aggregating Proteins. Side-chain hydrophobicity plays a well-established role in diseases involving aggregating proteins (53–55). To test whether the trends observed in Fig. 2*D* are amplified in aggregating proteins, we separated aggregating proteins from the primary dataset. Proteins involved in the formation of extracellular amyloid deposits or intracellular inclusions with amyloid-like characteristics are included in the "aggregating proteins" subset (28 proteins, 124 nSNPs, 330 dSNPs), while all remaining proteins are labeled as "nonaggregating proteins."

First, we used whole-sequence blobulation ($H^\star = 0.4$, $L_{\min} = 4$) to compare the enrichment distributions for aggregating and nonaggregating proteins using two different approaches for blob characterization. In addition to the hydrophobicity class, blobs were also assigned charge class, which is the predicted globular

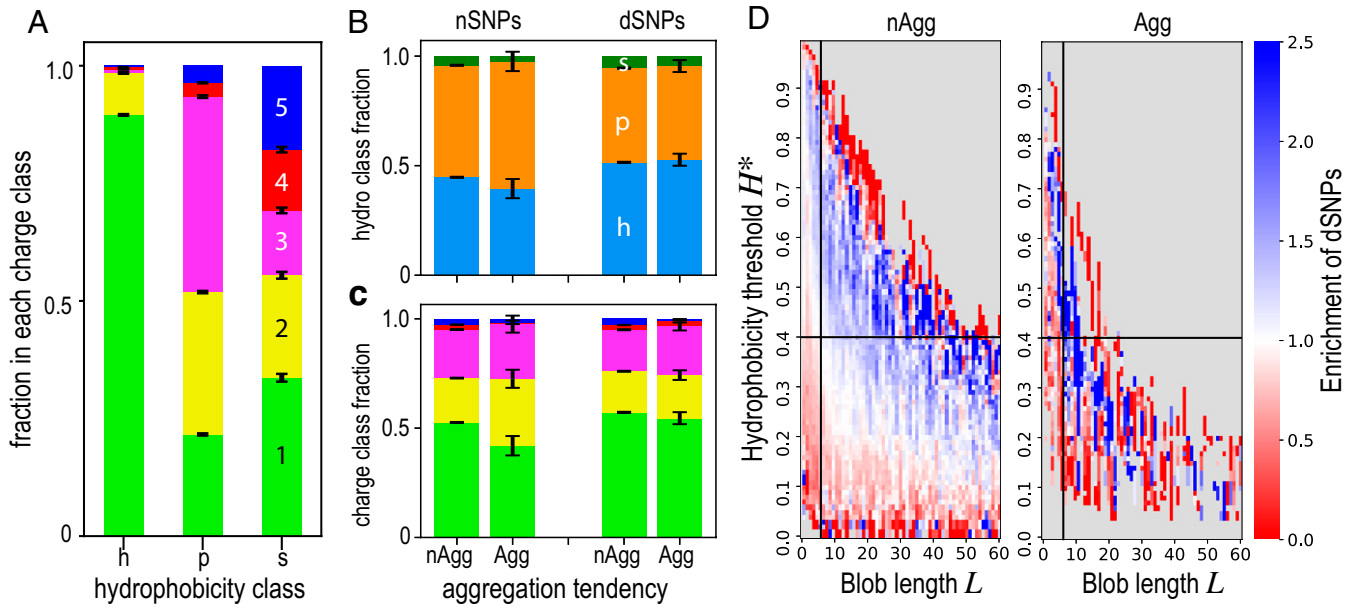

Fig. 4. Blob hydrophobicity and charge properties for SNPs in aggregating proteins. (*A*) Distribution of charge classes [Das–Pappu (56) phase] across the SNP dataset, for each blob hydrophobicity class. Possible values of the blob charge class are 1 (weak polyampholyte), 2 (Janus or boundary region), 3 (strong polyampholyte), 4 (negatively charged strong polyelectrolyte), and 5 (positively charged strong polyelectrolyte). (*B* and *C*) Fraction of nSNPs or dSNPs that are found in blobs of each hydrophobicity class (*B*) or charge class (*C*) in nonaggregating (nAgg) proteins and known-aggregating (Agg) proteins. *A*–*C* used whole sequence blobulation ($H^\star = 0.4$, $L_{\min} = 4$); error bars represent one SE for multinomial distributed data. Division of nAgg and Agg proteins is described in *Materials and Methods*. (*D*) Enrichment of dSNPs binned by threshold $H^\star$ and length *L* of the SNP-containing blob, calculated using unconstrained-length blobulation as in Fig. 2*D*, but separated by aggregation tendency. Black lines are to guide the eye.

phase according to Das and Pappu (56). This property is calculated using the fraction of positive and negative charges in a blob and is particularly relevant for IDPs. Nearly all structured proteins fall in the class 1 (weak polyampholyte) part of the Das–Pappu phase diagram. In contrast to structured proteins, IDPs can be found in all five Das–Pappu phases, including class 2 (Janus or boundary region), class 3 (strong polyampholyte), class 4 (negatively charged strong polyelectrolyte), and class 5 (positively charged strong polyelectrolyte).

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. 4*A*, which breaks down the fraction of h- and p-blobs that fall in each Das–Pappu charge class. As expected, most h-blobs (90%) fall in class 1 (weak polyampholyte), followed by 9% in class 2 (Janus). The p-blobs are more evenly distributed across classes, with the highest fraction (42%) classified as strong polyelectrolytes. Fig. 4 *B* and *C* shows the SNP distributions for blobs with different hydrophobicity class and charge class, respectively. Since there are five charge classes and only three hydrophobicity classes, we hypothesized that in nonaggregating proteins, blob charge class would have a stronger association with disease than blob hydrophobicity class. Instead, we found that the strongest dSNP enrichment as a function of charge class (1.09-fold, $P < 10^{-58}$ for weak polyampholyte blobs) is comparable to or even slightly less than the strongest enrichment found for hydrophobicity class (1.15-fold, $P < 10^{-100}$ for h-blobs).

Furthermore, the charge-based and hydrophobicity-based classification schemes yield similar trends with protein aggregation: Fig. 4*B* shows that a given dSNP in an aggregating protein is just as likely to be found in an h-blob as if it were in a nonaggregating protein, and Fig. 4*C* shows an analogous result for dSNPs in various charge classes. However, nSNPs in aggregating proteins are found slightly less frequently in blobs classified as h-blobs or class 1 (globular, weak-polyampholyte) blobs than nSNPs in nonaggregating proteins. As a result, we do observe a small increase in overall dSNP enrichment for h-blobs/weak-polyampholyte blobs of aggregating proteins relative to nonaggregating proteins.

We then used the results from unconstrained-length blobulation to further stratify the dSNP enrichment in aggregating proteins by hydrophobicity threshold $H^\star$ and blob length. The enrichment calculations from Fig. 2*D* were partitioned between aggregating and nonaggregating proteins and are shown in Fig. 4*D*. The highly enriched (blue) band is shifted toward the origin for aggregating proteins, indicating that sensitivity to mutation is found in shorter and more weakly hydrophobic blobs. The differential enrichment values using whole-sequence blobulation in Fig. 4 *B* and *C* arise from collapsing the distributions in Fig. 4*D* along a single value of $H^\star$. This result suggests that lower-hydrophobicity thresholds may be appropriate for predicting disease risk in known-aggregating proteins and underscores the importance of a multidimensional analysis for distinguishing between different groups of proteins.

Disease-Associated SNPs Are Enriched for Mutations That Change Local Blob Characteristics and Overall Protein Blob Topology. Whole-sequence blobulation yields a series of h-blobs, connected by p- and s-blobs, which we term the "blobular topology." Such a topology is analogous to the classic protein topology of secondary structure elements, although the location of edges and number of elements may be distinct. A SNP can alter the blobular topology by moving a short stretch of contiguous residues above or below the minimum blob size, either forming a new small h-blob or dissolving an existing small h-blob, respectively. A SNP can also split a long h-blob by interrupting a long contiguous hydrophobic sequence or merge two smaller h-blobs into one long h-blob by removing such an interruption.

Here we tested whether the dSNPs were more likely to change the topology determined by whole-sequence blobulation ($H^\star = 0.4$, $L_{\min} = 4$). Fig. 5*A* displays the fraction of nSNPs and dSNPs that cause each type of topological change. In the background case we expect to see more formation than dissolution, since the

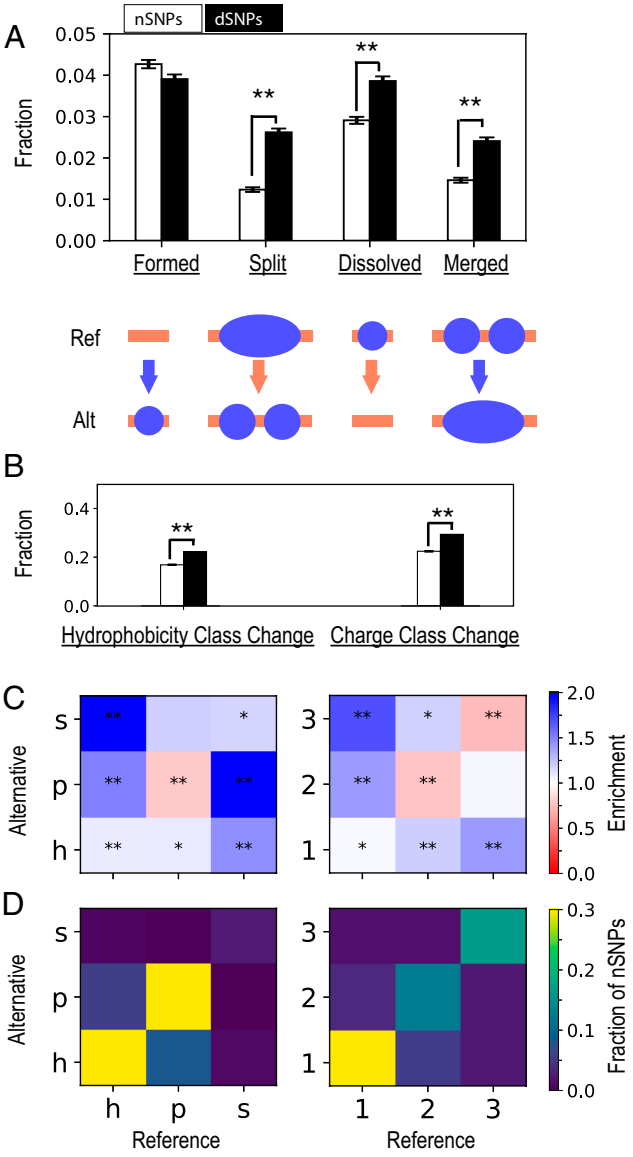

Fig. 5. Distribution of SNPs that change blob properties. (*A*) Fraction of nSNPs or dSNPs that change the blobular topology by either forming or dissolving an h-blob or by splitting one h-blob or merging two h-blobs. (*B*) Fractions of nSNPs or dSNPs that induce a change in hydrophobicity class or charge class. (*C*) The enrichment of dSNPs relative to nSNPs that induce a specific transition between the blob containing the reference allele (*x* axis) and the blob containing the alternative allele (*y* axis). This is shown for two blob properties, hydrophobicity class (*Left*) and charge class (*Right*), where the charge class categories are the same as in Fig. 4. (*D*) The overall proportion of nSNPs that induce each of the blobular topology transitions shown in *C*. The plot displays only charge classes 1 to 3 because fewer than 1.5% of SNPs involve charge class 4 or class 5. Significant enrichment or depletion in dSNPs is annotated with * ($P < 5 \times 10^{-3}$) or ** ($P < 5 \times 10^{-11}$) (binomial test). Errors bars in *A* and *B* represent one SE in the mean. All panels use whole sequence blobulation with $H^\star = 0.4$ and $L_{\min} = 4$.

blob count decreases with length (Fig. 2*H*), and there are more blobs just below the minimum length than just above it. Fig. 5*A* confirms this expectation for nSNPs, and the difference between the fraction of nSNPs and dSNPs that form new h-blobs is not significant (binomial test, $P > 10^{-3}$).

Other topological changes, however, are strongly enriched in dSNPs. As shown in Fig. 5*A*, dSNPs are significantly more likely to dissolve h-blobs (binomial test, $P < 10^{-21}$), split h-blobs (binomial test, $P < 10^{-80}$), and merge h-blobs (binomial test, $P < 10^{-36}$). 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 Fig. 2*D*. Regardless of overall topological changes, SNPs may change the characteristics of their local blob. Such changes may affect blob topology (as in Fig. 5*A*) or simply shift blob boundaries, causing a transition in blob class at the site of the SNP. The latter case is included in the data in Fig. 5*B*. We observe that about 17% of the nSNPs and 22% of dSNPs introduce a blob-hydrophobicity class change at the site of the SNP, yielding a 1.3-fold enrichment (binomial test, $P < 10^{-10}$; Fig. 5*B*). Fig. 5*C* compares the rates of specific blob hydrophobicity class transitions. Mutations involving h- $\rightarrow$ p-blob transitions yield the maximum enrichment (1.5-fold, $P < 10^{-10}$, binomial test) among dSNPs. In contrast, dSNPs are only 1.1-fold enriched (binomial test, $P < 10^{-3}$) in the reverse p- $\rightarrow$ h-blob transition, and SNPs that remain in p-blobs for both the reference and alternative allele are depleted among disease-associated SNPs.

Similarly, the frequency of SNP-induced changes in blob charge class is shown in Fig. 5*C*, for transitions between blobs in class 1 (weak polyampholyte), class 2 (Janus), or class 3 (strong polyampholyte). Transitions involving class 4 or class 5 (positively or negatively charged strong polyelectrolytes) represented fewer than 1.5% of the total transitions. Disease-associated SNPs are enriched for all mutations that change blob charge class and are either unenriched or weakly depleted for mutations that do not change the local blob charge class. Collectively, these results are consistent with the increased mutational sensitivity of hydrophobic (and typically buried) blobs that is shown in Fig. 2*D*, 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 charged residue was in a weak 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 peptide sequence and tested its use in detecting functional modules across the proteome. We show that hydrophobic h-blobs in solvated proteins are more likely to be buried: The h-blob SASA is 60% that of non–h-blobs. H-blobs are also nearly twice as likely to contain β-strands, supporting their proposed role as tertiary interaction sites. We find that enrichment of disease-associated mutations in hydrophobic blobs increases with the strictness of the hydrophobic blob criteria, with greater than fourfold enrichment for disease association in the longest, most hydrophobic blobs. This result persists when SNPs in transmembrane domains are removed from the analysis. The range and resolution of varying enrichment are 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 that is tied to blob hydrophobicity. Combined, these observations support our hypothesis that blobulation provides a more meaningful and less noisy approach to protein segmentation than use of a fixed-length moving window.

We also find that disease-associated mutations are significantly more likely than non–disease-associated mutations to change the blob topology of the sequence. This suggests that blobulation provides a meaningful topology that can be used as a framework for sequence analysis and requires only the protein amino acid sequence and two parameters (minimum blob length and hydrophobicity threshold). Once blobs are identified, they can be characterized using any property of interest. As an example, in the present work, we 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 (53), ProA (54), and Zyggregator (55). We do find that functional sensitivity in aggregating proteins follows similar trends to those in nonaggregating proteins, 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.

We demonstrated straightforward use of blobulation in combination with another residue characterization method like the Das–Pappu charge class. In this usage, blobulation serves only as the underlying segmentation approach for defining the local sequence. In principle, secondary structure prediction could be used for defining the local sequence instead. As demonstrated by the example of ubiquitin (Fig. 1*E*), however, secondary structure elements can cross between different faces of the protein, so secondary structure boundaries may not capture functional boundaries. Many secondary structure prediction methods require alignment to a homologous sequence with known structure (57–60), which may not be available. Yet this information is essential for secondary structure predictors to achieve their primary goal: determining which secondary structure the segment will adopt. Predicting segmentation, however, requires only determining the segment boundaries. Sequence-informed segmentation is a more feasible and straightforward task than structure prediction and yet has been largely unexplored in existing sequence analysis methods.

Improved information about the local context of a SNP, particularly those in hydrophobic blobs, could aid the identification of functional mutations. In the context of interpreting genome-wide association study results (61), the blob characteristics surrounding an associated variant provide metrics for fine mapping and ranking putative causal variants. Such blob metrics could also be used as input features for predicting variant function with machine-learning algorithms, which derive their decision rules based on training datasets of annotated mutations. To this end, we provide a two-dimensional table (Dataset S4) of disease-association enrichment as a function of blob properties.

Many of the analyses in this paper use the relaxed criteria for h-blobs to demonstrate that even conservative stratification yields meaningful differences. Thus, the potential applications are not limited to those residues that meet the strict criteria or have the strongest disease association. We view the parameter sensitivity of

blobulation as a methodological strength, because adjusting the two parameters allows the user to "zoom" in or out by tuning the number of detected edges. In our parallel development of a blobulation graphical interface (see *Materials and Methods* for access information), we have repeatedly observed that such tuning can bring previously obscured sequence organization into visual focus.

Materials and Methods

Blobulation. The algorithm is illustrated in Fig. 1 *A* and *B*. As shown in Fig. 1*A*, for a given peptide, every amino acid i is assigned a mean hydrophobicity H_i, defined as the average Kyte–Dolittle (33) hydropathy score with a window size of three residues, scaled to fit between 0 and 1. The sequence is then digitized by testing whether $H_i > H^\star$ for each amino acid; if $H_i > H^\star$, then residue i is classified as hydrophobic, and if not, residue i is classified as nonhydrophobic. Note that the algorithm classification is solely dependent on the Kyte–Dolittle score and the threshold $H^\star$, rather than the canonical classification of residue types. For instance, serine and threonine are not canonically hydrophobic, but typically have a hydropathy score beyond the relaxed threshold $H^\star = 0.4$. Even charged residues can be classified as "hydrophobic" if they are surrounded by hydrophobic residues (so that $H_i > 0$) and the threshold $H^\star$ is sufficiently low.

After the sequence is digitized, the sequence is blobulated, as shown in Fig. 1*B*. The algorithm first identifies all contiguous stretches of at least $L_{\min}$ hydrophobic residues; these stretches are classified as h-blobs. Of the remaining subsequences in the given peptide, those that are at least as long as $L_{\min}$ are termed p-blobs, while those shorter than $L_{\min}$ are termed s-blobs. Example effects of varying $H^\star$ and $L_{\min}$ are shown in Fig. 1*C*. The underlying software engine and a web interface for sequence analysis with adjustable parameters are freely available as described in *Computational Packages*.

In addition to the classic whole-sequence blobulation method just described, we also use two variants that fix a reference residue i and relax either of the two parameters. Unconstrained-length blobulation fixes the threshold $H^\star$ and calculates the length L of the blob containing residue i, rather than imposing a minimum length. Similarly, unconstrained-threshold blobulation calculates $H_{\max}$, which is the maximum possible value of $H^\star$ that would still assign residue i to an h-blob that meets the fixed $L_{\min}$ requirement on minimum blob length.

Secondary and Tertiary Structural Analysis. We used the Ensembl BioMart tool (62) to select human proteins that also had available structures in the PDB. Only one structure was chosen for each unique sequence; for those sequences with multiple available structures, we used a structure with maximum residue coverage. In total, the structural dataset contained 6,459 proteins, each of which were blobulated with $L_{\min} = 4$ and $H^\star = 0.4$. The secondary structure for each residue of a given blob type was calculated using the DSSP algorithm (63, 64). For the secondary structure calculations shown in Fig. 1*C*, "helix" consists of alpha-helix, 3-helix, and 5-helix; "beta" consists of isolated beta bridge and extended strand; and "coil" consists of all the remaining DSSP secondary structure types, including turn and bend. Transmembrane domains (identified using UniProt annotations) (36) were removed from this dataset for the SASA calculations. For each residue, the raw SASA value was calculated using DSSP and then divided by the residue-specific maximal accessibility (65) to determine the relative SASA. The relative SASAs were then averaged for all residues of a given blob type. The SASA data are provided in Dataset S1 and the secondary structure data in Dataset S2.

SNP Datasets. The SNP data we use is the UniProtKB literature-curated list of missense variants (https://www.uniprot.org/docs/humsavar, obtained on 17 June 2020) (66). Variants are annotated using the American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) terminology (67). dSNPs are those annotated as "likely pathogenic or pathogenic" ($N = 30{,}227$), and nSNPs are those annotated as "likely benign or benign" ($N = 39{,}448$). SNPs in transmembrane domains were identified using the annotations in UniProt (36).

dSNP Enrichment Tests. For a given residue annotation, such as being in an h-blob, we test whether there are proportionally more dSNPs with that annotation than expected based on the proportion of nSNPs with this annotation. The enrichment we report is the ratio of the dSNP proportion over nSNP proportion. To quantify the statistical significance, we use a binomial test on the dSNP count assuming the nSNP proportions apply. Specifically, if we observe n dSNPs with a given annotation out of N total dSNPs, and if the proportion of nSNPs with this annotation is f, then under a null model the observed dSNPs count is a binomial experiment of N "tests," each with independent probability of "success" f. We compute the probability of observing a count as extreme as n (two-tailed) given f using the python scipy (68) function scipy.stats.binom_test(n,N,f,alternative= "two-sided").

Fixed-Length Moving Windows. For each SNP i, we compute the mean hydrophobicity $\langle H \rangle_i$ within a window of length L_w centered on i. For a given threshold $H^\star$, SNP i is classified as falling in a "hydrophobic window of length L_w" only if $\langle H \rangle_i \geq H^\star$. The window lengths L_w were iterated over all odd numbers between 1 and 99 (or the protein sequence length if the protein was less than 99 residues long), so that equal numbers of residues were included on each side of the SNP. The enrichment of dSNPs in hydrophobic windows compared to nSNPs is calculated as described in *dSNP Enrichment Tests*.

Population Frequency Data. Frequency data are from the gnomAD v3 genomes dataset of variants for the non-Finnish European cohort: gnomad.genomes.r3.0.sites.chr*.vcf.bgz, accessed 24 July 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 apply the blobulation algorithm to every SNP i to find the maximum blob hydrophobicity threshold, $H_{\max}$, for which at least one of the alleles remains in an h-blob with length $\geq L_{\min} = 4$. SNPs for which neither allele resides in an h-blob of minimum length $L_{\min}$, regardless of $H^\star$, are discarded from further analysis. We blobulate in increments of $\Delta H^* = 0.05$ from $H^\star = 0$ to $H^\star = 1$ to identify $H_{\max}$ with a resolution of 0.05. 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 sets. The preceding procedure is tabulated for nSNPs and dSNPs separately. The frequency data for the SNPs used in the analysis for Fig. 3 are provided in Dataset S5.

Gene Pathway and Ontology Enrichment Tests. The gene-ontology enrichment is performed using the g:Profiler web service (69). 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 Encylopedia 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 (HPO). The above results are provided in Dataset S6.

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, and P02766. For analyses involving aggregation, the UniProt dataset was divided into the SNPs within these 28 proteins (aggregating proteins) and all other proteins (nonaggregating proteins).

Das–Pappu Charge Class. The blob charge class is a secondary blob property representing the Das–Pappu phase (56), which is determined using the fraction of positively and negatively charged residues as originally prescribed. Blobulation does not rely on charge class, but any blob may be assigned a charge class following blobulation.

Blob Transitions Induced by a SNP. Whole-sequence blobulation ($H^\star = 0.4$, $L_{\min} = 4$) is performed on two protein sequences, each containing either the reference or the alternate allele. The reference allele is used for all residues except *i*, even if the protein contains multiple other SNPs. Topological changes are identified via scanning the two sequences. The hydrophobicity class and charge class of the SNP-containing blob are also determined for each sequence. The proportion of dSNPs that induce a specific transition (or no transition) is tested for enrichment as described in *dSNP Enrichment Tests*, where the residue-level annotation is the specific transition caused by the SNP.

Computational Packages. All computations were done in Python 3.6 using the numpy (70), scipy (68), and pandas (71) packages. All plots are made using the Python matplotlib (72) package. Molecular images were made using Visual Molecular Dynamics (VMD) (73).

Data Availability. All data used in this study are available in Datasets S1–S6, with detailed legends for each given in the *SI Appendix*. Briefly, Datasets S1 and S2 cover the structural, SASA, and blobulation data used for Figure 1. Dataset S3 is the unconstrained-length blobulation data per SNP that constitutes the base SNP dataset used for the blobulation data in Figures 2, 3, 4, and 5. Dataset S4 contains the enrichment data per bin shown in Figure 2. Dataset S5 contains the SNP frequency data used in Figure 3. Dataset S6 contains the gene pathway enrichment results. The public repository for the blobulation software (including a script for reading output from the webtool into VMD for visualization) is available (74). At the time of publication, a graphical webtool for blobulation of user-provided sequences is also available (75).

ACKNOWLEDGMENTS. We acknowledge the Office of Advanced Research Computing 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. M.E.B.H. is supported by NIH Grants 1R35GM134957, R01AR076241, and ADA 1-19-VSN-02, and G.B. is supported by the Busch Biomedical Foundation. We are also grateful to Ms. Kaitlin Bassi and Mr. Connor Pitman for useful discussions and comments on the manuscript, as well as to all other past and current members of the Blobulator development team (Ms. Kaitlin Bassi, Ms. Lindsey Riggs, Mr. Connor Pitman, Mr. Ezry Santiago-McRae, Dr. Thomas Joseph). Finally, we are grateful to Dr. Cameron Abrams for coining the term "blobulation" to describe our algorithm.

Author affiliations: [a]Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102; [b]Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724; [c]Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104; and [d]Department of Physics, Rutgers University, Camden, NJ 08102

1. J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, D. T. Jones, Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635–645 (2004).
2. H. J. Dyson, P. E. Wright, Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208 (2005).
3. V. N. Uversky, Unusual biophysics of intrinsically disordered proteins. *Biochim. Biophys. Acta.* **1834**, 932–951 (2013).
4. V. N. Uversky, Intrinsically disordered proteins and their "mysterious" (meta)physics. *Front. Phys.* **7**, 10 (2019).
5. A. R. Panchenko, M. M. Babu, Editorial overview: Linking protein sequence and structural changes to function in the era of next-generation sequencing. *Curr. Opin. Struct. Biol.* **32**, viii–x (2015).
6. P. H. Weinreb, W. Zhen, A. W. Poon, K. A. Conway, P. T. Lansbury Jr., NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. *Biochemistry* **35**, 13709–13715 (1996).
7. V. N. Uversky, L. M. Iakoucheva, A. K. Dunker Protein disorder and human genetic disease. *eLS*, 10.1002/9780470015902.a0023589 (2012).
8. K. Cuanalo-Contreras, A. Mukherjee, C. Soto, Role of protein misfolding and proteostasis deficiency in protein misfolding diseases and aging. *Int. J. Cell Biol.* **2013**, 638083 (2013).
9. A. Patel *et al.*, A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell* **162**, 1066–1077 (2015).
10. V. N. Uversky, Intrinsically disordered proteins and their (disordered) proteomes in neurodegenerative disorders. *Front. Aging Neurosci.* **7**, 18 (2015).
11. L. Tovo-Rodrigues *et al.*, The role of protein intrinsic disorder in major psychiatric disorders. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **171**, 848–860 (2016).
12. R. Lohia, R. Salari, G. Brannigan, Sequence specificity despite intrinsic disorder: How a disease-associated Val/Met polymorphism rearranges tertiary interactions in a long disordered protein. *PLOS Comput. Biol.* **15**, e1007390 (2019).
13. E. A. Stone, A. Sidow, Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.* **15**, 978–986 (2005).
14. A. Niroula, S. Urolagin, M. Vihinen, PON-P2: Prediction method for fast and reliable identification of harmful variants. *PLoS One* **10**, e0117380 (2015).
15. V. López-Ferrando, A. Gazzo, X. de la Cruz, M. Orozco, J. L. Gelpí, PMut: A web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Res.* **45** (W1), W222–W228 (2017).
16. M. Hecht, Y. Bromberg, B. Rost, Better prediction of functional effects for sequence variants. *BMC Genomics* **16** (suppl. 8), S1 (2015).
17. P. Popov, I. Bizin, M. Gromiha, K. A, D. Frishman, Prediction of disease-associated mutations in the transmembrane regions of proteins with known 3D structure. *PLoS One* **14**, e0219452 (2019).
18. P. C. Ng, S. Henikoff, Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
19. P. D. Thomas, A. Kejariwal, Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: Evolutionary evidence for differences in molecular effects. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 15398–15403 (2004).
20. E. Capriotti, R. Calabrese, R. Casadio, Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* **22**, 2729–2734 (2006).
21. Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, A. P. Chan, Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**, e46688 (2012).
22. E. Capriotti, P. Fariselli, R. Casadio, A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics* **20** (suppl. 1), i63–i68 (2004).
23. E. Capriotti, P. Fariselli, R. Casadio, I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* **33**, W306–W310 (2005).
24. V. Parthiban, M. M. Gromiha, D. Schomburg, CUPSAT: Prediction of protein stability upon point mutations. *Nucleic Acids Res.* **34**, W239–W242 (2006).
25. G. Wainreb *et al.*, MuD: An interactive web server for the prediction of non-neutral substitutions using protein structural data. *Nucleic Acids Res.* **38**, W523–W528 (2010).
26. S. Ittisoponpisan *et al.*, Can predicted protein 3d structures provide reliable insights into whether missense variants are disease associated? *J. Mol. Biol.* **431**, 2197–2212 (2019).
27. S. Iqbal *et al.*, Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 28201–28211 (2020).
28. A. Prlić *et al.*, Integrating genomic information with protein sequence and 3D atomic level structure at the RCSB protein data bank. *Bioinformatics* **32**, 3833–3835 (2016).
29. P. W. Rose *et al.*, The RCSB protein data bank: Integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* **45** (D1), D271–D281 (2017).
30. S. Mir *et al.*, PDBe: Towards reusable data delivery infrastructure at protein data bank in Europe. *Nucleic Acids Res.* **46** (D1), D486–D492 (2018).
31. L. Jiang *et al.*, Real-time monitoring of hydrophobic aggregation reveals a critical role of cooperativity in hydrophobic effect. *Nat. Commun.* **8**, 15639 (2017).
32. L. Lins, A. Thomas, R. Brasseur, Analysis of accessible surface of residues in proteins. *Protein Sci.* **12**, 1406–1417 (2003).
33. J. Kyte, R. F. Doolittle, A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
34. B. C. Finzel, T. L. Poulos, J. Kraut, Crystal structure of yeast cytochrome c peroxidase refined at 1.7-A resolution. *J. Biol. Chem.* **259**, 13027–13036 (1984).
35. S. Gao *et al.*, Monomer/oligomer quasi-racemic protein crystallography. *J. Am. Chem. Soc.* **138**, 14497–14502 (2016).
36. UniProt Consortium, UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49** (D1), D480–D489 (2021).
37. S. Teng, A. K. Srivastava, L. Wang, Sequence feature-based prediction of protein stability changes upon amino acid substitutions. *BMC Genomics* **11** (suppl. 2), S5 (2010).
38. E. Capriotti, P. Fariselli, PhD-SNPg: A webserver and lightweight tool for scoring single nucleotide variants. *Nucleic Acids Res.* **45** (W1), W247–W252 (2017).
39. K. Chen, L. Kurgan, J. Ruan, "Optimization of the sliding window size for protein structure prediction" in *2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology* (IEEE, 2006).
40. A. Schlessinger, B. Rost, Protein flexibility and rigidity predicted from sequence. *Proteins* **61**, 115–126 (2005).
41. O. Sander, I. Sommer, T. Lengauer, Local protein structure prediction using discriminative models. *BMC Bioinformatics* **7**, 14 (2006).
42. Y. Park, S. Hayat, V. Helms, Prediction of the burial status of transmembrane residues of helical membrane proteins. *BMC Bioinformatics* **8**, 302 (2007).
43. C. Baeza-Delgado, M. A. Marti-Renom, I. Mingarro, Structure-based statistical analysis of transmembrane helices. *Eur. Biophys. J.* **42**, 199–207 (2013).
44. R. E. Dickerson, The structures of cytochrome c and the rates of molecular evolution. *J. Mol. Evol.* **1**, 26–45 (1971).
45. B. Charlesworth, M. T. Morgan, D. Charlesworth, The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1993).
46. S. R. Sunyaev, W. C. Lathe III, V. E. Ramensky, P. Bork, SNP frequencies in human genes an excess of rare alleles and differing modes of selection. *Trends Genet.* **16**, 335–337 (2000).
47. A. Siepel *et al.*, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
48. M. Somel *et al.*, A scan for human-specific relaxation of negative selection reveals unexpected polymorphism in proteasome genes. *Mol. Biol. Evol.* **30**, 1808–1815 (2013).
49. V. Llaurens, A. Whibley, M. Joron, Genetic architecture and balancing selection: The life and death of differentiated variants. *Mol. Ecol.* **26**, 2430–2448 (2017).
50. K. J. Karczewski *et al.*, Genome Aggregation Database Consortium, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

51. S. Alonso, S. López, N. Izagirre, C. de la Rúa, Overdominance in the human genome and olfactory receptor activity. *Mol. Biol. Evol.* **25**, 997–1001 (2008).
52. D. Pierron, N. G. Cortés, T. Letellier, L. I. Grossman, Current relaxation of selection on the human genome: Tolerance of deleterious mutations on olfactory receptors. *Mol. Phylogenet. Evol.* **66**, 558–564 (2013).
53. O. Conchillo-Solé *et al.*, AGGRESCAN: A server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics* **8**, 65 (2007).
54. Y. Fang, S. Gao, D. Tai, C. R. Middaugh, J. Fang, Identification of properties important to protein aggregation using feature selection. *BMC Bioinformatics* **14**, 314 (2013).
55. G. G. Tartaglia, M. Vendruscolo, The Zyggregator method for predicting protein aggregation propensities. *Chem. Soc. Rev.* **37**, 1395–1401 (2008).
56. R. K. Das, R. V. Pappu, Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 13392–13397 (2013).
57. Y. Yang *et al.*, Sixty-five years of the long march in protein secondary structure prediction: The final stretch? *Brief. Bioinformatics* **19**, 482–494 (2018).
58. B. Zhang, J. Li, Q. Lü, Prediction of 8-state protein secondary structures by a novel deep learning architecture. *BMC Bioinformatics* **19**, 293 (2018).
59. Y. Wang, H. Mao, Z. Yi, Protein secondary structure prediction by using deep learning method. *Knowl. Base. Syst.* **118**, 115–123 (2017).
60. Y. Ma, Y. Liu, J. Cheng, Protein secondary structure prediction based on data partition and semi-random subspace method. *Sci. Rep.* **8**, 9856 (2018).
61. M. D. Gallagher, A. S. Chen-Plotkin, The post-GWAS era: From association to function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).
62. K. L. Howe *et al.*, Ensembl 2021. *Nucleic Acids Res.* **49** (D1), D884–D891 (2021).
63. W. Kabsch, C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
64. W. G. Touw *et al.*, A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* **43**, D364–D368 (2015).
65. S. Miller, J. Janin, A. M. Lesk, C. Chothia, Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641–656 (1987).
66. Y. L. Yip *et al.*, Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. *Hum. Mutat.* **29**, 361–366 (2008).
67. S. Richards *et al.*, ACMG Laboratory Quality Assurance Committee, Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
68. P. Virtanen *et al.*, SciPy 1.0 Contributors, SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
69. U. Raudvere *et al.*, g:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47** (W1), W191–W198 (2019).
70. C. R. Harris *et al.*, Array programming with NumPy. *Nature* **585**, 357–362 (2020).
71. The Pandas Development Team, pandas-dev/pandas: Pandas. https://doi.org/10.5281/zenodo.3509134. Accessed 15 October 2021.
72. J. D. Hunter, Matplotlib: A 2d graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
73. W. Humphrey, A. Dalke, K. Schulten, VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38, 27–28 (1996).
74. Blobulator Development Team, Protein Blobulator. GitHub. https://github.com/BranniganLab/blobulator. Deposited 3 September 2022.
75. Blobulator Development Team, Blobulator. https://www.blobulator.branniganlab.org. Accessed 23 January 2022.