Comprehensive Analysis of the Spatial Distribution of Missense Variants in Protein Structures Reveals Patterns Predictive of Pathogenicity

By

Robert Michael Sivley

Thesis
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE
in

Biomedical Informatics
May, 2017
Nashville, Tennessee

Approved:
John A. Capra, Ph.D.
William S. Bush, Ph.D.
Jens Meiler, Ph.D.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Motivation

To determine the genetic etiology of disease, we must understand the functional effects of disease-causing genetic variants. Causal variants that alter amino acids in the protein-coding sequences of genes (i.e., missense variants) are believed to predominantly derive their pathogenicity from the alteration of protein structure, and consequently, protein function. Through evolutionary and molecular analyses, we know that functionality is not distributed evenly throughout a protein. Few amino acids in a protein structure compose active sites or binding interfaces; most amino acids are responsible for protein folding and stability. The mechanism by which a missense variant does or does not disrupt protein function is dependent on the part of the protein it affects and the degree to which the specific amino acid substitution affects it. While the specific function of many proteins is not always well characterized, evolutionary conservation analyses can identify which amino acids in a protein have been highly conserved across species, indicating functional importance. Initial evidence from a small number of proteins suggests that these evolutionarily conserved amino acids are spatially clustered within protein structures. If this pattern holds for all proteins, the spatial analysis of evolutionarily conserved amino acids in protein structure has the potential to accurately resolve regions of functional importance. While sequence conservation highlights the importance of some amino acids, clear spatial boundaries can inform functional hypotheses about why they are important.

If functionally important residues are spatially clustered, we hypothesize that functionally disruptive, disease-causing missense variants will also be spatially clustered. While evolutionary conservation may identify amino acids (or regions) of functional importance, it provides no information about what function is being performed. Disease-causing variation can often infer the functional role of a region from the phenotypic outcome of its disruption. With the recent abundance of whole-genome and whole-exome sequencing data, coupled with growing numbers of experimentally derived and computationally predicted protein structures, we have an opportunity to investigate these patterns on the scale of the human proteome.

Evidence for Conserved Amino Acid Clustering in Protein Structure

Identifying functional regions of protein structures is currently difficult, as our understanding of protein function is incomplete. Computational modeling and prediction is steadily improving with the advent of methods like ddg_monomer[1] and VIPUR[2], but at present we cannot reliably predict (in an accurate, high-throughput manner) the structural impact of amino acid substitutions or how the resulting structural changes will ultimately influence function without expert analysis. While it is sometimes possible to empirically determine the functional effect of a given substitution, such assays are inherently low-throughput and limited to proteins for which a functional assay is available. Sequence conservation can help to bridge this gap by identifying amino acids that have been highly conserved throughout evolution, suggesting they play an important role in the function or stability of that protein.

This measure of functional importance has motivated several efforts towards the analysis of evolutionary conservation in protein structure. ConSurf-DB[3] was designed to identify and

visualize the most and least conserved amino acids in a protein structure, independent from the protein's baseline degree of conservation. Capra *et al.* found that the combination of evolutionary sequence conservation and protein structure significantly improved detection of ligand-binding sites over both conservation-only and structure-only methods.[4] Madabushi *et al.*[5] more directly evaluated the spatial clustering hypothesis, calculating Evolutionary Trace[6] for each amino acid in 46 proteins from different structural and functional classes. They found that conserved amino acid clusters were significantly larger than expected (relative to a random distribution) in 45 of the 46 proteins analyzed. The utility of this spatial clustering pattern has also been recognized in the field of protein fold prediction. Baker *et al.* [7] identified significant clustering of conserved residues in 73 of 79 analyzed protein structures. They then demonstrated that algorithmic constraints requiring conserved residues to be in close spatial proximity dramatically improve *de novo* protein structure prediction. Other structural analyses of evolutionary conservation include the identification of conserved positions in protein folds,[8] binding interfaces,[9] and the prediction of functional sites.[10–13] Each of these analyses is limited in either scope or scale, but the consistency of their findings supports the hypothesis that evolutionarily conserved residues cluster in protein structures and that those clusters represent functional regions of proteins.

Evolutionary Conservation is Predictive of Variant Pathogenicity

Disease-causing missense variants are presumed to derive their pathogenicity from the disruption of protein function and stability. As discussed above, functional amino acids can be identified through evolutionary conservation analysis. It follows that missense variants affecting

evolutionarily conserved amino acids are more likely to disrupt protein function. Application of this hypothesis has been overwhelmingly successful in variant pathogenicity prediction. All nine of the popular pathogenicity prediction algorithms evaluated by Thusberg *et al.*[14] incorporate some measure of evolutionary conservation[15–22].

Mendelian Germline Variants and Recurrent Somatic Mutations Cluster in Protein Structure

The strong evidence for conserved residue clustering and the relationship between evolutionary conservation and variant pathogenicity suggests that disease-causing missense variants also form clusters in protein structure and that these clusters identify functional regions of proteins with relevance to specific diseases. Indeed, the literature already includes initial evidence for the pathogenic clustering of Mendelian and somatic cancer mutations. In an analysis of 162 diseases affecting 181 genes, Turner *et al.*[23] found that both dominant and recessive disease-causing mutations were significantly more clustered than neutral variants from the 1000 Genomes Project[24]. Numerous studies have also analyzed the clustering of somatic mutations in cancer[25–31], where discriminating between driver and passenger mutations is a challenge. Despite differences in methodology, all have identified significant clustering of somatic mutations in both the sequence and structure of oncogenes, tumor suppressors, and genes not previously associated with cancer. These germline and somatic analyses are analogous. Oncogene and dominantly inherited mutations are largely presumed to result in gain-of-function, while tumor suppressor and recessive mutations are presumed to result in loss-of-function. It is likely that these analyses are detecting a similar phenomenon: pathogenic variants in close spatial proximity share similar functional and disease-related properties. This phenotypic similarity is highlighted

by *Guo et al*. in their comprehensive analysis of three-dimensional protein interactome networks. They found that recessive mutations in interacting proteins were significantly more likely to cause the same disease if they affected the binding interface of those two proteins[32].

Current Methods for Quantifying Missense Variant Clustering

There is substantial evidence to support the hypothesis that pathogenic variants form spatial clusters in protein structure, but the methods for evaluating these distributions are redefined by each subsequent publication and are often focused towards the identification of a few known examples. In this section I will discuss four methods for quantifying the degree of spatial clustering observed for missense variants along with their strengths and weaknesses.

A straightforward way of evaluating clustering amongst a set of points is the sum of inverse pairwise distances. Stehr *et al*.[30] adopted this approach in their analysis of somatic mutations in 24 oncogenes and tumor suppressors (Figure 1A). Within each protein structure, they calculated the sum of inverse pairwise distances amongst somatic mutations from COSMIC[33] and population-derived missense single-nucleotide polymorphisms (SNPs) from the 1000 Genomes Project. They conclude that all missense variants, regardless of pathogenicity, were significantly more clustered than expected at random. Moreover, somatic mutations in oncogenes were more significantly more clustered than population-derived SNPs, while somatic mutations in tumor suppressors were not. Although the sample size is small and limited to known cancer genes, the results of this analysis defined expectation for future studies of the spatial distributions of somatic mutations.

While the sum of inverse pairwise distances is simple to compute, it incorporates no information about biologically meaningful distances or the frequency at which a mutation is observed in individuals with cancer. Kamburov *et al.*[26] addressed these concerns with the introduction the two transformations shown in Figure 1B. First, each variant is weighted according to its frequency using somatic mutation counts from The Cancer Genome Atlas (TCGA)[34]. The mutation frequency is transformed, such that mutations observed in only one individual receive values near 0, while mutations observed in six or more individuals receive a value of 1. Each pair-wise distance is then transformed such that highly proximal mutations receive values near 1, and distal variants receive values approaching 0. The extent of clustering within the protein structure is then scored as the sum of the weighted, pairwise distances between recurrence-weighted somatic mutations. Ultimately, proteins with highly recurrent mutations in close spatial proximity receive high scores and those containing distal mutations with low recurrence receive low scores. The significance of these scores is evaluated by randomly permuting the locations of somatic mutations within the protein structure. In their analysis of 4,062 human proteins, 10 were found to harbor significantly clustered recurrent somatic mutations at a false discovery rate (FDR) $q<0.1$. Unlike Stehr *et al.*, Kamburov *et al.* observe significant clustering in both oncogenes and tumor suppressors, as well as in proteins not previously associated with cancer; they attribute this difference largely to methodological differences and dataset selection. A weakness of this approach is the reliance on user-intuition in defining biologically plausible distances and recurrence counts. Additionally, neither of these approaches identify discrete clusters of mutations nor do they define functional boundaries; rather, they report that somatic mutations within the protein structure are more clustered than the null expectation.

In contrast, the approach proposed by Meyer *et al.*[25] aims to define clusters directly and then evaluates the significance of the clusters themselves. It accomplishes this with complete-linkage hierarchical clustering using the distance between somatic mutations (Figure 1C). Complete-linkage clustering enforces a maximum distance between any two mutations in the same cluster. This distance is user-defined, but a value of 15Å was used for their comprehensive analyses. The algorithm additionally enforces the minimum cluster size (number of mutations) of three or more unique amino acid substitutions at two or more unique protein positions. Justification of these parameters is not provided. The significance of each cluster was determined through random permutation of the mutations within the protein structure. Clusters that contained more mutations than expected at random were considered significant. Although this approach defines clear mutation clusters, there are also several weaknesses. First, many of the algorithm's parameters are user-defined without clear evidence for biologically plausible arguments. Second, many of the clusters identified by the algorithm are heavily influenced by highly recurrent mutations at only two protein positions. While these recurrent mutations may disrupt the same function, it is unlikely that cluster analysis is required to identify their importance.

Most recently, a novel cluster analysis was reported by Tokheim *et al.*[27] that incorporates many of the attributes from the above work. The first stage of this analysis evaluates the local missense mutation density around each somatic mutation. This is defined as the sum of the count of missense mutations observed at each mutation position and the count of missense mutations within 10Å of that site (Figure 1D). A distance threshold of 10Å was chosen because it is the order of magnitude of an amino acid side chain[27]. Permutation of the mutation positions within the structure is used to generate an empirical null and determine which mutations are within "mutation hotspot regions." Once identified, these mutations are grouped into clusters using a

neighbor graph such that neighbors within 10Å of one another are connected. Each subgraph is then defined as a mutation cluster. With this approach, Tokheim *et al*. can define clusters of variable shape and size while still restricting local density measures to a biologically plausible scale. However, the method still makes assumptions about what distance scales are biologically plausible. Also, because no recurrence normalization was applied to the somatic mutation counts (as described in Kamburov *et al*.) and because the mutation count for a given protein position is included in its local mutation density estimate, the method tend to identify individual (or pairs of) highly recurrent mutations that are likely identifiable without cluster analysis.

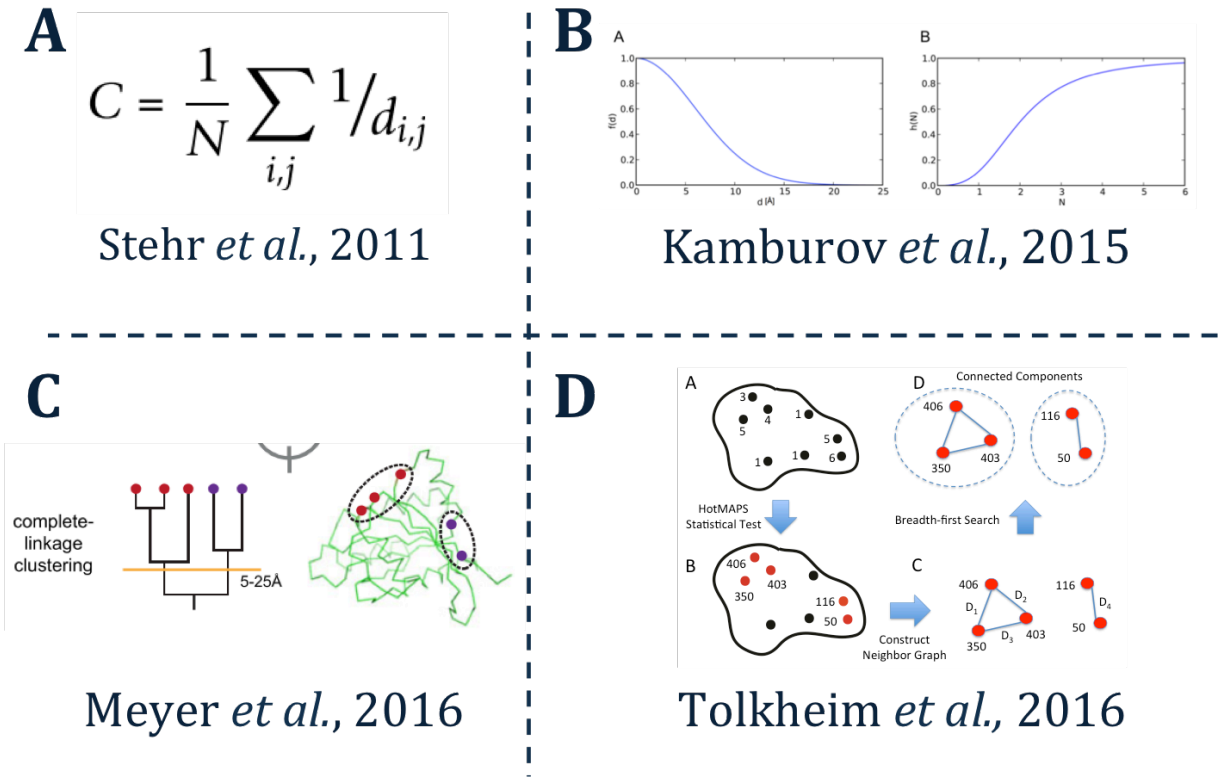

Figure 1: Methodological illustrations of approaches for identifying missense variant clustering in protein structure.

Limitations of Hypothesis-Driven Clustering Methodologies

Each of the methods described above was designed specifically for the identification of clusters of somatic mutations, with heavy influence from known examples in cancer. This limits the generalizability and applicability of these methods for testing alternative spatial hypotheses. Collectively, these methods have several limitations. The first is a reliance on experimenter intuition. All but Stehr *et al.* include intuition-based parameterization: Kamburov *et al*. transform distance and recurrence, Meyer *et al.* impose an arbitrary upper bound on cluster diameter and lower bound on mutation counts within observed clusters, and Tokheim *et al.* define distance bounds for the "local" neighborhood of a mutation that are based on the approximate length of an amino acid side chain. Within each study, justification was provided for why these normalizations and bounds are necessary, but little justification was provided for their parameterization. While it is important that method parameters and results are within biologically plausible limits, each of these decisions narrows the types of clusters that can be discovered. For example, a major benefit of hierarchical clustering is its ability to detect non-spherical clusters, but the complete-linkage aggregation and maximum distance threshold used by Meyer *et al.* will lead the algorithm towards small, spherical clusters and cannot accurately recognize non-spherical clusters if the longest axis exceeds 15Å. These restrictions limit its effectiveness in analyzing transmembrane proteins or protein-protein binding interfaces where functional regions may be non-spherical. It is also unclear if parameters inferred from somatic mutation clusters will be optimal for other types of genetic variation and structural properties. There has been significant, recent interest in identifying genes with significantly less population-derived missense variation than expected by chance[35]. Evaluating the regional patterns of this

phenomenon within protein structure would require a fundamentally different hypothesis about the expected and observed spatial distribution of (putatively neutral) missense variants. Each of the methodologies described above would require significant modification to evaluate these new hypotheses because they've been designed specifically to identify clusters of disease-causing variants.

The second major limitation is the narrow scope of previous work. Most of these analyses focus heavily on clusters of somatic mutation in cancer, but the functional properties driving missense spatial constraint should be similarly applicable to germline missense variation. Meyer *et al*. briefly discuss an increased likelihood of clustering Human Gene Mutation Database (HGMD)[36] pathogenic variants relative to putatively benign variants from the Exome Variant Server (http://evs.gs.washington.edu/EVS/), but that analysis pools together pathogenic and putatively benign variants to simulate the mixture of driver and passenger mutations in cancer. Ultimately, the comparison is only discussed as justification for analyzing somatic mutation data. Some analyses have instead focused on the spatial clustering observed for evolutionary conservation, but these studies have been limited in scale. Despite heavy reliance on the conclusions of several small studies, there has not been a comprehensive, systematic evaluation of conserved residue clustering across all human proteins.

Many previous studies have compared the general patterns of clustering observed for somatic mutation with the patterns observed for population-derived missense variants, but none provide the option of evaluating one dataset in relation to another. For example, Stehr *et al.* found that somatic mutations in tumor suppressors were no more clustered than population-derived missense variants. This conclusion was drawn from a comparison of the global trends observed for the two datasets. A more direct hypothesis for this comparison would test whether somatic

mutations were more or less clustered than population-derived missense variants within each structure. This more specific hypothesis may identify specific proteins in which the clustering of somatic mutations is significantly more than for population-derived missense variation, in contrast to the global trend.

The methodology for analyzing spatial distributions in protein structure should not be influenced by domain-specific hypotheses. A general, data-driven approach is needed that is equally capable of analyzing evolutionary and structural properties as well as germline and somatic protein-coding variation. This approach should then be applied to the numerous datasets of evolutionary, genetic, and structural data publically available to evaluate previously observed patterns and novel hypotheses on a comprehensive scale. These data are not independent from one another, but are largely influenced by similar biological pressures. This approach should enable comparisons between synonymous and missense variants, conserved amino acids and pathogenic variants, or germline and somatic variants. In the following chapters, I describe such an approach along with a pipeline and database to support the high-throughput, spatial analysis of evolutionary conservation and human genetic variation within protein structure.

Chapters

The process of linking genetic variant information to amino acid coordinates in experimentally derived protein structures is a non-trivial task involving numerous resources and databases, each with their own format, cross-references, and inconsistencies. In chapter 2, I provide a detailed description of the PDBMap pipeline and database. The purpose of this resource is the linkage of

existing databases of genetic and structural information to facilitate efficient evaluation of structural hypotheses about missense variation on a global scale.

Once missense variants have been placed within protein structure, we require a robust statistical framework to quantify their spatial distribution. In chapter 3, I describe a general methodology for evaluating spatial distributions within and between datasets of protein-coding variation and structural properties. These methods make no assumptions about cluster size, shape, or scale, are sensitive to both clustering and dispersion, and are not tailored to find domain-specific examples. With this framework, we aim to quantify on a large-scale whether evolutionarily conserved amino acids are clustered in protein structures. We next evaluate whether spatial patterns of missense variation are derived from the effects of amino acid substitution by contrasting synonymous and nonsynonymous distributions. With these properties defined, we evaluate the hypotheses that pathogenic missense variants are clustered in protein structures and that the degree of clustering exceeds what is observed for neutral variants. Finally, we highlight the utility of spatial information by evaluating its predictive performance in classifying pathogenic and neutral missense variants.

CHAPTER II

PDBMAP: MAPPING PROTEIN-CODING VARIATION INTO PROTEIN STRUCTURE

Introduction

Although the processes of gene transcription, mRNA translation, and protein folding are biologically linked, the scientific fragmentation of genetics, proteomics, and structural biology introduces technical barriers to holistic analysis. These limitations impede large-scale analysis of genetic variation in the context of protein structure and complicate the evaluation of otherwise testable hypotheses. In this chapter, I present PDBMap, a pipeline and database for explicit mapping between the human genome and structome. The initial pipeline pre-computes the mapping between all amino acids in all experimentally derived and computationally predicted protein structures and all protein-coding nucleotides in the human genome. Using this resource, genetic variants and annotations can be directly mapped into all available protein structures without navigating the entire cross-reference network. Finally, we populate this database with the largest public datasets of protein structures and genetic variation: The Protein Data Bank[37] (27,624 human protein structures), ModBase[38] (102,235 human homology models), ExAC[35] (4,521,130 synonymous and missense variants), and ClinVar[39] (56,162 missense variants).

Protein Structural Databases

The Protein Databank

The Protein Data Bank (PDB) is the central repository for experimentally derived protein structures with over 5,000 distinct human proteins represented by more than 27,000 structures. Because of the many-to-many relationship of proteins to structures, there are analyses included in this project that limit the dataset to a representative subset of individual protein chains. This subset was generated by Kamburov *et al.*[26] and is intended to provide the greatest coverage of the human proteome while minimizing redundancy. Although protein structures in the PDB are determined by a variety of experimental methods, all those included in this subset were derived from either X-ray crystallography or solution NMR. All solution NMR structures were represented by the first recorded model.

ModBase

ModBase is a public repository of comparative protein structure models. Their automated pipeline uses the Modeller[40] software to generate homology models for the database. Homology modeling involves the identification of proteins with high sequence identity to a protein of interest, which is likely to adopt a similar tertiary fold. The homologous structure is then used as a template along which the sequence of the target protein is threaded. From this starting state, a series of loop-building and relaxation processes attempt to identify the lowest energy conformation, which is presumed to be the protein's native state. The quality of computationally predicted homology models is highly variable and largely dependent on sequence identity with the template structure. Models are evaluated using the ModPipe quality score (MPQS), a

composite score incorporating sequence identity of the target protein with the identified template structure, structural coverage of the target protein, and three external quality scores: e-value, z-Dope, and GA341. ModBase considers a MPQS > 1.1 to indicate a reliable structural model. The ModBase database is included to supplement the Protein Data Bank and improve coverage of proteins without experimentally derived protein structures to over 17,000 distinct human proteins.

Genetic Variation Databases

The Exome Aggregation Consortium (ExAC)

Whole-exome sequencing provides a comprehensive view of protein-coding variation within the human genome. To elucidate the mechanisms behind disease-causing variation, we must first understand the patterns of the neutral variation we all carry in abundance. ExAC aggregates whole-exome sequencing data from fourteen cohorts totaling 60,706 individuals and six human continental populations: African/African American, Latino, East Asian, South Asian, Finnish, and Non-Finnish European. This resource provides a high-resolution view of where putatively neutral variation is tolerated within human proteins and serves as a neutral background for the analysis of pathogenic missense variation.

ClinVar

Managed by the National Center for Bioinformatics (NCBI), ClinVar is a submission-based database of variant-disease associations. Clinical significance is assigned in accordance with recommendations by the American College of Medical Genetics (ACMG), which stratifies

variant disease associations into five classes: uncertain significance, benign, likely benign, likely pathogenic, and pathogenic. ClinVar represents our current understanding of the genetic cause of many complex diseases, but also includes Mendelian associations reported by the Online Mendelian Inheritance in Man (OMIM).

Linking Genetic Variants to Protein Structure

There are three primary components required to link genetic and protein structural information. The first involves the reconciliation of protein structural sequences with reference protein sequences and alignment with the human genome. The second is consequence determination for protein-coding genetic variants. Third, these two sources must be joined using shared identifiers to create a complete path from genetics to protein structure, enabling high-throughput analyses. A schematic of the PDBMap pipeline and database is provided for reference in Figure 1.

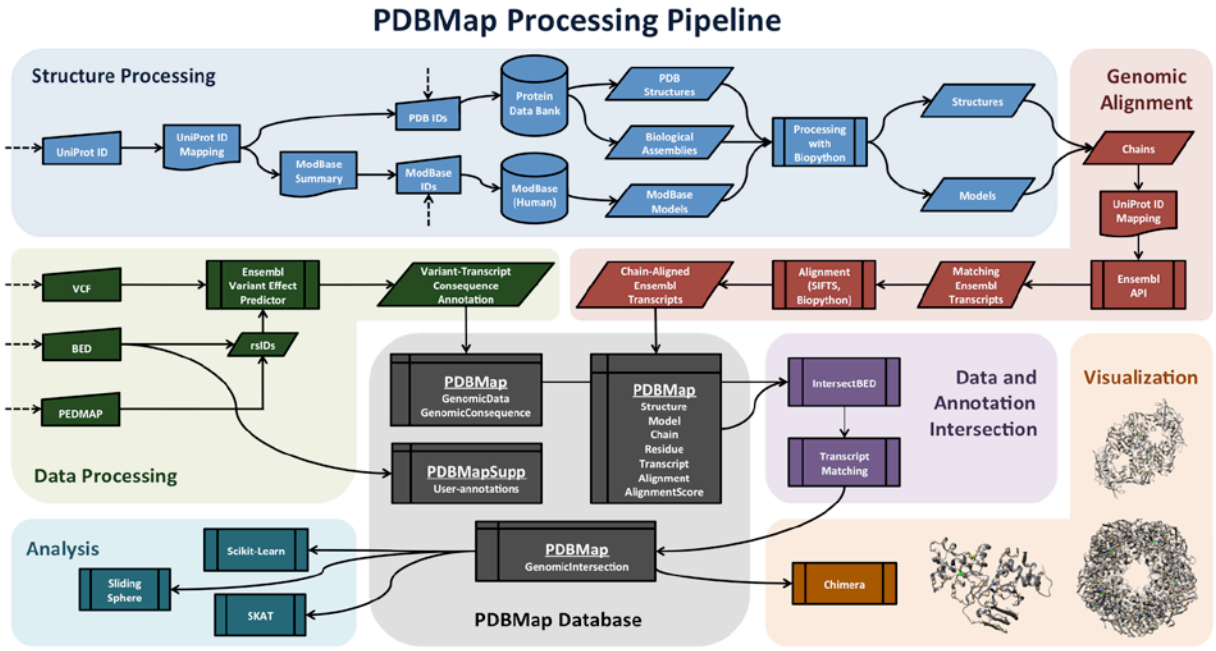

Figure 2: A schematic overview of the PDBMap pipeline and database. Structural processing tasks are shown in blue. Alignment of protein chains with gene transcripts is shown in red. Consequence prediction for genetic variation is shown in green. Intersection of genetic datasets with the structure-to-genome mapping is shown in purple. Light blue and orange describe downstream applications of this resource. Dashed arrows indicate where external data is fed into the pipeline.

Aligning protein structures to the human genome

A protein structure is a snapshot of the three-dimensional conformation assumed by a human protein. However, the experimental process of deriving these structures is a source of both intentional (e.g. expression tags, mutations) and incidental (e.g. alternative residue numbering) deviations from the reference protein sequence. In many cases, these errors can be resolved through simple pairwise alignment, but recently a more robust alternative was released. SIFTS[41] is a manually curated database of protein structural information that, among other useful annotations, provides a residue-level mapping from structure to reference sequence for each amino acid in a protein. Further, protein sequences not matching the reference sequence (e.g.

expression tags, mutations) can be easily identified and/or removed. Any remaining alignment errors are corrected through simple pairwise alignment.

Having aligned each chain of each protein structure to its reference protein sequence (structome to proteome), we next attempt to match UniProt proteins with EnsEMBL gene transcripts (proteome to genome). For each protein accession identifier, UniProt provides a list of corresponding EnsEMBL gene transcripts. However, the two databases often disagree, primarily due to differences in the handling of genes with multiple isoforms and homologous genes that produce highly similar transcripts. This issue is compounded by the lack of protein isoform information in the PDB. This introduces a many-to-many relationship between gene transcripts and protein isoforms, where not all transcript-isoform matches are valid. We again approach this step as an alignment problem. Rather than aligning protein structure sequences to reference protein sequences, we instead align them directly to the translated peptide sequences of EnsEMBL gene transcripts. Alignments with high sequence identity (>90% with respect to protein structure) are considered valid matches (not only does the protein identifier match, but the specific protein isoform matches the sequence encoded by the gene transcript). This approach identifies and removes transcripts that could not plausibly produce the sequence observed in the protein structure. The 90% threshold is intended to accommodate minor alterations to the protein sequence, like point mutations or short gaps in the protein structure.

Using the aligned EnsEMBL transcripts for each protein structure, we derive the nucleotide coordinates for each amino acid in the protein sequence. On completion of this pipeline, every amino acid in every protein structure is directly annotated with both its structural and genomic coordinates. Of the 20,160 human proteins reviewed by Swiss-Prot, PDBMap currently includes

27,624 protein structures containing 5,223 (26%) distinct human proteins and 102,235 homology models containing 17,341 (86%) distinct human proteins.

Determining variant consequence

Not all protein-coding variants cause a change in the protein sequence. Some may introduce stop-codons that truncate the protein, while some nucleotide substitutions produce codons that encode the same amino acid. Similarly, a single variant may have different consequences or affect different amino acid positions for different gene transcripts. We determine variant consequences using EnsEMBL's Variant Effect Predictor (VEP)[42] tool. In addition to determining transcript-specific consequences for each genetic variant, VEP also provides numerous informative annotations including global and population allele frequencies, gene and transcript annotations, and pathogenicity predictions from established tools like SIFT and PolyPhen2. None of the information reported by VEP is required for structural mapping (only the genomic position is required), but consequence prediction is useful for validation and critical for understanding the potential effects of each variant on the protein structure.

Intersecting genetics and structure

We've presented above the structural and genetic components of the PDBMap database. The first provides a direct link between the structural and genomic coordinates of each amino acid in a protein structure. The second determines the consequence of each genetic variant for each transcript of the affected gene. Both datasets now contain a collection of shared identifiers that we can use to efficiently map all genetic variants into protein structure. This alignment is conducted in two stages.

To create the initial mapping from genetic variant to protein structure, we generate two sets of genomic coordinates. The first is the set of genomic ranges that define the codon for each amino acid in a protein structure. The second is a set of genomic positions (or ranges, if INDELS are present) identifying the location of each genetic variant. Using intersectBED[43], we can efficiently intersect these two datasets to identify all variant positions that overlap codons mapped to amino acids in a protein structure, in effect mapping genetic variants into protein structure. We then use the consequence predictions from VEP to eliminate matches where the transcript associated with the variant consequence does not match the transcript aligned to the protein sequence. Finally, we verify that the affected protein position predicted by VEP matches the reference position in the matched protein structure. These final two steps use information derived from different sources to eliminate erroneous mappings. Upon completion of the intersection and validation, a direct link is created between each genetic variant and each amino acid in every protein structure affected by that variant. The current status of genetic datasets in the PDBMap database is provided in Table 1.

	Total	**Mapped to PDB**	**Mapped to ModBase**
ExAC Synonymous	1,549,333	178,626 (12%)	955,173 (62%)
ExAC Missense	2,971,797	300,604 (10%)	1,809,201 (61%)
ClinVar Missense	56,162	13,997 (25%)	38,894 (69%)
COSMIC Missense	1,366,383	158,985 (12%)	881,446 (65%)

Table 1: Number of distinct synonymous or missense variants within each genetic dataset.

High-throughput analysis of genetic variation in protein structure

The alignment of a large whole-exome sequencing dataset with all protein structures in the Protein Data Bank and/or ModBase is a complicated and time-consuming task. Running the described pipeline for a dataset like ExAC, parallelized across chromosomes, requires just over a day of processing time. However once the process is complete and the results uploaded to the PDBMap database, the coordinates of any variant or set of variants in any or all affected protein structures can be rapidly determined with a simple MySQL query. For example, querying the structural coordinates of all ExAC synonymous and missense variants in all solved protein structures requires just over one second of processing (plus download time).

Conclusion

This resource enables high-throughput, large-scale analysis of protein-coding variation within protein structures. In the following chapters, we present several spatial analyses comparing different classifications of protein-coding, single-nucleotide variation. All of the work included in those chapters is derived from and dependent upon the structure-to-genome mapping provided by the PDBMap database.

CHAPTER III

QUANTIFYING SPATIAL PATTERNS OF GERMLINE PROTEIN-CODING VARIATION IN PROTEIN STRUCTURES

Introduction

The foundational work presented in Chapter II facilitates efficient, large-scale analysis of missense variation in its protein structural context. As described in Chapter I, several approaches have been used in the analysis of somatic mutations in cancer. These analyses have largely been focused towards specific hypotheses and the re-identification of canonical examples from cancer. In this chapter, we present a general, statistical framework – adapted from Ripley's K[44–46] – for defining and evaluating the spatial distribution of missense variants (and other residue-level annotations) in protein structures. We focus the application of this methodology towards the analysis of evolutionary conservation and germline variation. We first evaluate the fundamental hypothesis that missense variants are spatially constrained and that this constraint is not observed in synonymous variation. We next determine if the clustering of evolutionarily conserved amino acids is a general phenomenon. Finally, we evaluate the spatial distributions of disease-causing missense variation, contrast it with neutral variation, and evaluate the predictive performance of these spatial relationships.

Methods

Protein structure selection, variant mapping, and annotation

Three single-nucleotide variant (SNV) datasets were included in our comprehensive analyses: Exome Aggregation Consortium[35] (ExAC) r0.3, and ClinVar (01-07-2016). Synonymous SNVs in ExAC were included for comparison with ExAC missense SNVs. All other datasets were reduced to missense SNVs. Variant consequences and annotations were determined using v82 of the Ensembl Variant Effect Predictor.[42] Additional dominant and recessive HGMD[36] missense variants from Turner *et al.*[23] were used to investigate gain- and loss-of-function spatial patterns.

Ensembl[47] transcripts were matched with UniProt[48] accession and Protein Data Bank[37] (PDB) IDs using ID-mapping tables provided by UniProt. Reference protein sequences were aligned with observed sequences in the PDB using SIFTS.[41] Any discrepancies were corrected by pairwise alignment with Biopython.[49] Proteins were represented by the subset of minimally overlapping PDB structures described by Kamburov *et al.*[26]. For each protein, the algorithm selects the PDB chain with the greatest coverage of the protein sequence. This process continues iteratively, excluding PDB chains with greater than 10% sequence overlap with the set of already selected chains, until the complete sequence is structurally represented or all available structures have been processed. Evolutionary conservation was calculated by Jensen-Shannon divergence[50] using multiple sequence alignments from HSSP[51].

For each missense variant in ExAC, we identified the Euclidean nearest neighbor using genomic position and protein structural coordinates. Genomic nearest neighbors were restricted to other missense variants within the same gene. Structural distances were measured from the

centroid of each amino acid side-chain. Missense variants without a neighbor in either the gene or protein structure were excluded from the comparison.

Ripley's K for quantifying spatial distributions in protein structure

Ripley's *K* is a test for spatial heterogeneity that measures the deviation of a set of positions from complete spatial randomness (CSR), capturing both clustering and dispersion. Because missense variants are constrained to the positions of amino acids in a protein structure, the assumption of CSR for randomly distributed variants is inappropriate. In a constrained space, permutation testing provides an empirical null distribution for comparison with observed patterns. At each distance threshold, the number of neighbors around each variant is compared with the empirical null expectation. When the number of neighbors exceeds expectation, the variants are clustered; when the number of neighbors is lower than expectation, the variants are dispersed. Because *K* is measured across a range of distances, it is possible to identify clustering or dispersion at any scale. Our estimator for *K* is defined as,

$$\widehat{K} = \frac{\sum_{i}^{N} \sum_{j!=i}^{N} I(D_{ij} < t)}{N(N-1)}$$

where N is the number of variants, D_{ij} is the Euclidean distance between variants *i* and *j* in the protein structure, and *I* is an indicator function that evaluates to 1 when D_{ij} is less than the distance threshold *t* and 0 otherwise. *N(N-1)* is applied as a normalization factor such that *K* is the proportion of variant pairs within distance *t*. This normalization allows for comparison between proteins with different variant counts. Variant positions are defined as the centroid of the reference amino acid. Permutations are generated by randomly sampling *N* amino acids from the protein structure over 100,000 iterations and recomputing *K*. Two-tailed p-values are derived from the percentile rank of the observed *K* value relative to the distribution of permuted *K*

values. Z-scores are also calculated to indicate the direction (clustering or dispersion) and magnitude of the effect. An illustrative description of the *K* analysis is shown in Figure 1.

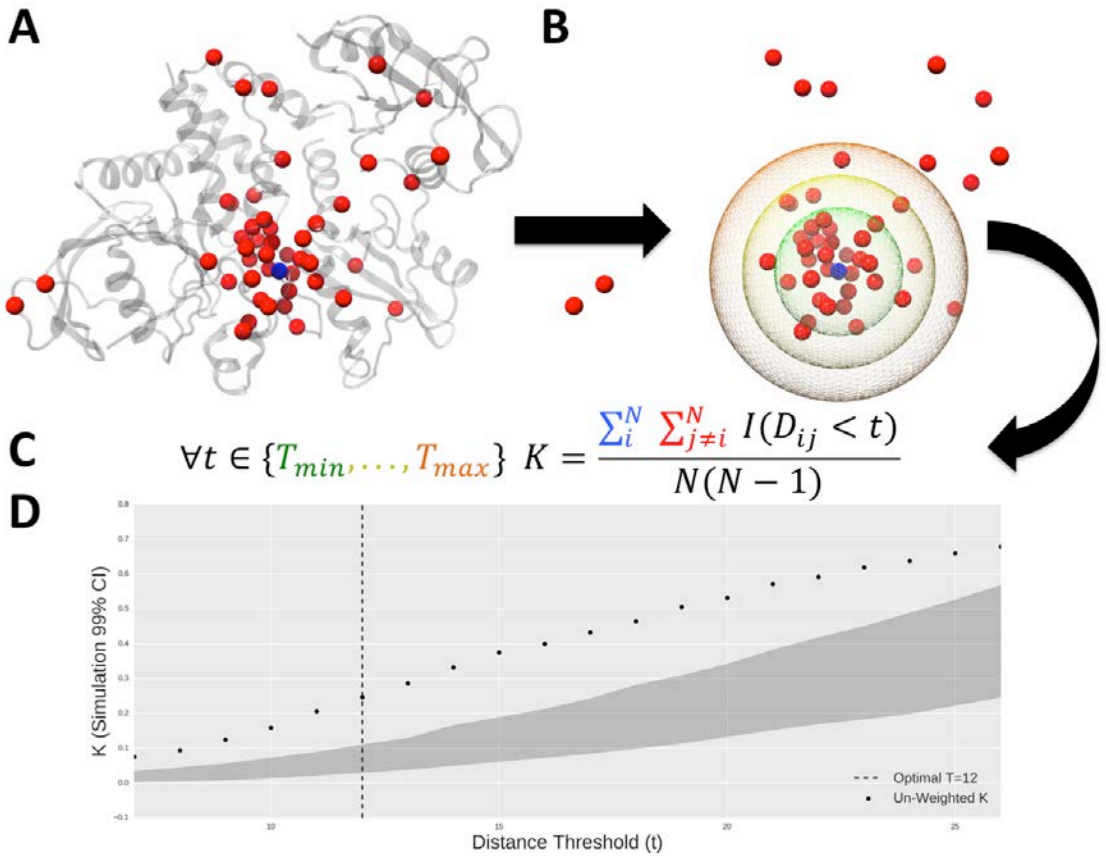

Figure 3: Quantifying the spatial distribution of missense variants in protein structure with Ripley's K. (A) Missense variants are mapped into solved structures from the Protein Data Bank. (B) Using a range of distance thresholds, (C) the proportion of variants within that radius is measured. (D) An empirical null distribution is determined through random shuffling of variant positions (un-weighted univariate) or variant labels (bivariate and weighted univariate) and used to calculate permutation p-values.

Ripley's *K* can also evaluate real-valued data to determine if the values are spatially correlated, conditioned on the variant positions. The weighted *K* is define as,

$$\widehat{K}_{weighted} = \frac{\sum_i^N \sum_{j!=i}^N I(D_{ij} < t) w_j}{\sum_i^N \sum_{j!=i}^N w_j}$$

where w_j is the weight applied to variant j. A variant cannot be counted as its own neighbor, so the weight of w_i is not included in the function. The implication of this restriction is that a single, high-valued variant cannot form its own cluster.

The weighed *K* analyzes the process by which weights, not positions, are distributed. Thus, the significance of the weighted *K* is calculated by permuting weights over fixed positions and recomputing *K*. This approach assesses whether the assigned values are significantly clustered or dispersed beyond what is explained by variant position. When *N* is small, our ability to assess the significance of the weighted *K* is constrained by the number of unique permutations of the sample weights.

Ripley's *K* is designed to quantify spatial distributions across a range of distances, capturing clustering or dispersion at any scale. To evaluate spatial patterns only at biologically relevant distance scales, we define the distance range from the minimum observed distance between variants to half the maximum observed distance between variants. Proteins for which the minimum distance was greater than half the maximum distance were extended to the maximum distance.

While the ability to investigate spatial patterns at multiple distances is valuable, a protein-level summary statistic is required to compare between proteins and identify those containing significant spatial patterns. Each protein is summarized by the difference between observed K values and the median empirical null K values. The area between the curves is defined as the difference of their integrals, as estimated by Simpson's rule. This summarization captures the direction (clustering or dispersion) and magnitude (absolute z-score) of the multi-distance signal observed for that protein. The distance threshold yielding the most significant signal is also

retained to approximate the scale at which the spatial signal is strongest. Finally, a protein summary z-score and p-value is determined by calculating the summary *K* for each permutation, such that positive z-scores indicate clustering, negative z-scores indicate dispersion, and z-scores near zero indicate spatial randomness (e.g. a lack of spatial constraint). To control for a False Discovery Rate of 10%, we calculate q-values from the protein-summary p-value distribution in each analysis.

Bivariate K for spatial comparisons between variant datasets

The univariate *K* is useful for quantifying biases in the spatial distribution of a single dataset, but many biological questions involve comparisons between variants of different types (e.g. neutral and deleterious). These comparisons are best made with bivariate *K* functions. The simplest form of the bivariate test examines whether one set of positions is more or less clustered than another using the difference in *K* between the two datasets.

$$\widehat{D} = \widehat{K}_N - \widehat{K}_M$$

Similar to the weighted univariate *K*, the bivariate *D*[45] evaluates the process by which dataset labels are assigned. It follows that the significance of the bivariate *D* is determined through random permutation of the class labels over fixed positions.

Relative proximity to pathogenic variation as a predictor of pathogenicity

To measure the proximity of an unlabeled variant to a set of known variants, the distance between each variant is first transformed by the NeighborWeight function[52],

$$NeighborWeight(x, y, lower\ bound, upper\ bound) = \begin{cases} 1, if\ d_{x,y} \leq lower\ bound \\ \frac{1}{2}\left[\cos\left(\frac{d_{x,y} - lower\ bound}{upper\ bound - lower\ bound} \times \pi\right) + 1\right], \\ if\ lower\ bound < d_{x,y} < upper\ bound \\ 0, if\ d_{x,y} \geq upper\ bound \end{cases}$$

where $d_{x,y}$ is the distance between variants *x* and *y*. A lower bound of 8Å provides full weight to amino acids for which direct interaction is plausible. A lower bound of 24Å centers the cosine curve on 16Å, providing larger weights to variants with the potential for indirect interaction and smaller weights to variants with some likelihood of affecting similar structural regions or domains. Using this transformation, we define the average proximity of an unlabeled variant *x* to a set of variants *Y* as,

$$P_{x,Y} = \sum_{y}^{Y} \frac{NeighorWeight(x, y, 8{,}24)}{|Y|}$$

The pathogenic proximity score for each variant is then defined as the relative proximity to pathogenic and putatively neutral variation,

$$\Delta P_x = P_{x,pathogenic} - P_{x,neutral}$$

such that values of ΔP_x greater than 0 indicate that variant *x* is more nearby pathogenic variants than neutral variants. Using leave-one-out cross validation, we rank variants by their relative proximity to ClinVar pathogenic and ExAC missense variants and calculate receiver-operator-characteristic (ROC) and precision-recall (PR) curves. We evaluate predictive performance using area under ROC and PR curves (AUC). These results are then compared using Analysis of Variance (ANOVA) with pathogenicity predictions from PolyPhen2, SIFT, and evolutionary conservation.

Results

Most SNVs have a different nearest neighbor in sequence and structure

Nearly all analyses of protein-coding variation are based on the position of variants within the linear nucleotide or protein sequence. More recent analyses have begun grouping variants by gene[53] or predicted functional domains[54]. These partitions are problematic for spatial analyses because protein folding dramatically alters the spatial distribution of amino acids, bringing linearly distant residues into close structural proximity. To assess the extent of this phenomenon, we examined each ExAC missense variant and identified its nearest neighboring missense variant by genomic and structural coordinates. We found that nearest neighbors differed between sequence and structure for 60% of structure-mapped missense variants. While this metric doesn't quantify the full impact of protein folding on the spatial landscape of missense variation, it highlights the potential impact of incorporating structural information into spatially informed aggregate analyses of protein-coding variation like collapsing tests, burden tests, and SKAT[55], which may not otherwise be capturing the most relevant functional groups of missense variants.

Synonymous and missense variants have different spatial distributions

An important assumption in our spatial analyses is that patterns of missense variation are influenced by the functional and structural impact of amino acid substitution. However, the spatial patterns we observe can also arise from other biological and technical effects, like inconsistencies in mutation rate, sequencing coverage, and similar effects. To test this assumption and evaluate the extent to which these effects influence the observed spatial distributions, we compared the spatial distributions of synonymous variants and missense

variants from ExAC. Synonymous variants do not alter protein sequence, so their distribution in protein structure should be under no spatial constraint. Thus, all deviations from spatial randomness can be attributed to the genomic inconsistencies that affect both synonymous and missense variation. By evaluating the spatial distribution of missense variants in protein structure in reference to the spatial distribution of synonymous variants, we can identify the patterns of spatial constraint attributable only to amino acid substitution. For each protein, we calculated the univariate *K* for synonymous and nonsynonymous variants. As expected, we find that synonymous variants show little divergence from random spatial distributions (Figure 2), with only one protein reaching statistical significance (MYOM1, FDR<0.1). Conversely, missense variants display a general trend towards spatial dispersion and were significantly non-randomly distributed in 51 proteins; 37 with significant dispersion and 14 with significant clustering (Table 1). There was a highly significant difference between the synonymous ($N_{synonymous}$=4,498) and missense ($N_{missense}$=4,487) spatial distributions ($p=2.71x10^{-120}$ Mann Whitney U), supporting the hypothesis that missense variants are under increased spatial constraint relative to synonymous variants.

Despite the difference in the univariate trends, a bivariate analysis directly comparing synonymous and missense variants in 4,173 proteins identified only two in which missense variants were significantly more dispersed than synonymous. These results suggest that while globally, missense variants are consistently more dispersed than synonymous variants, within any given protein structure the sample size is likely too small and the difference too subtle to reach significance.

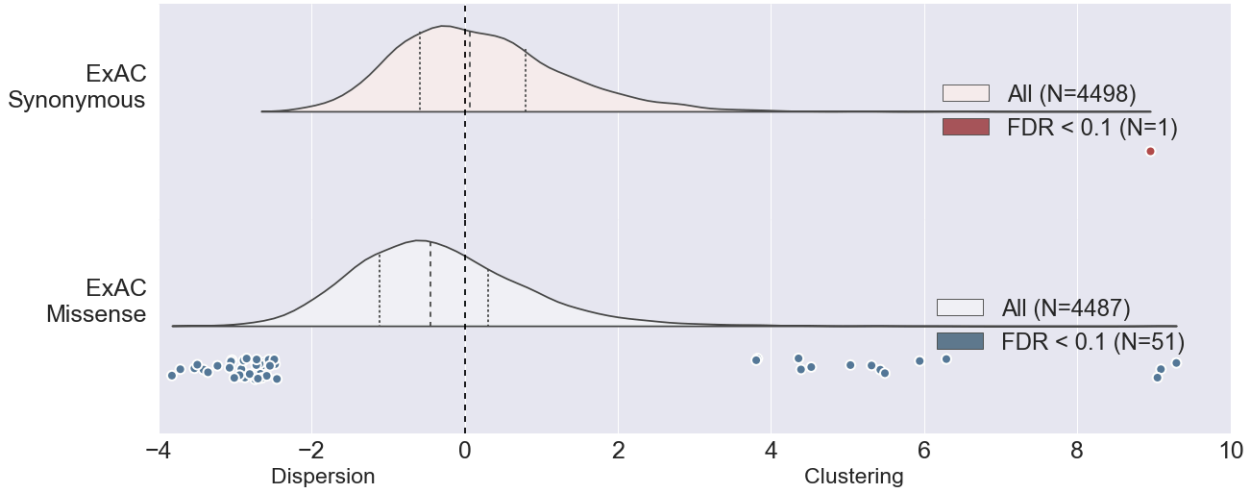

Figure 4: The distribution of protein-summary z-scores describes the general spatial patterns of a variant dataset. Synonymous variants from ExAC are generally randomly distributed, as indicated by a near-normal distribution of z-scores with median near 0. In contrast, missense variants from ExAC trend towards spatial dispersion, with significant spatial patterns identified for 51 proteins.

Spatial dispersion identifies a tendency for protein surface residues

In an unconstrained space, dispersion manifests as an even spacing between observations (Figure S1). Because protein structure is a constrained space, we hypothesized that dispersion would be greatest for amino acids at the protein surface. This is supported by a previously observed bias for population-derived missense variants to preferentially alter surface residues.[56] We performed two analyses to investigate whether signals of spatial dispersion identified by the univariate *K* were correlated with surface exposure.

We first performed a weighted, univariate K analysis of relative solvent accessibility (RSA). If collections of surface-exposed residues yield high dispersion values, then RSA should yield highly significant dispersion across all structures. Indeed, we observed significant spatial dispersion in 4,114 of 4,495 proteins (92%, FDR<10%) (Figure 3).

We next evaluated whether missense variants where more solvent accessible than all residues, which would indicate a bias for residues at the protein surface. We found that the RSA of all missense variants ($N_{all,missense}$=209,841) was significantly greater than the RSA of all residues ($N_{all,residue}$=972,121) ($p \approx 0$ Mann Whitney U) and that the RSA of missense variants in significantly dispersed missense variants ($N_{dispersed,missense}$=2,253) was significantly greater than the RSA of all missense variants ($p = 5.0 \times 10^{-54}$ Mann Whitney U). Interestingly, we also found that significantly clustered missense variants ($N_{clustered,missense}$=428) were no more or less solvent accessible than all residues ($N_{clustered,residue}$=2,902) (p=0.39 Mann Whitney U), suggesting that clusters of missense variants affect both the protein core and surface.

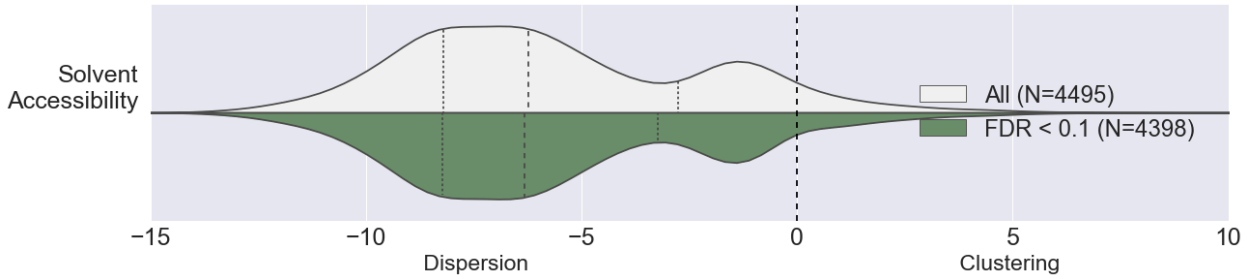

Figure 5: Distribution of protein z-scores for the weighted univariate analysis of relative solvent accessibility (RSA). The significant dispersion of RSA in 92% of proteins demonstrates that spatial dispersion identifies a bias for surface residues.

Evolutionary conservation is spatially constrained and generally clustered

As we currently lack a comprehensive understanding of the molecular function of all proteins in our study, evolutionary conservation serves as the most uniform measure of functional importance. Previous studies have demonstrated on a small scale that conserved residues form spatial clusters in protein structure[5] and that minimizing the distance between conserved residues can improve structure prediction[7]. Considering the conservation of sequence-adjacent residues

has also been shown to improve the identification of functionally important protein residues[50]. To evaluate the tendency for evolutionarily conserved residues to cluster in protein structure, we performed a weighted, univariate *K* analysis of evolutionary conservation scores. We identified significant clustering in 3,752 of 4,286 proteins (88%, FDR<0.1) and significant dispersion in 101 proteins (2%). (Figure 4). This finding confirms that clustering of evolutionarily conserved residues is a general phenomenon. Furthermore, it forms the basis for our analysis of pathogenic variation in protein structure, as it demonstrates that protein function – as measured by evolutionary conservation – is spatially constrained and detectable with our methodology. By directly analyzing pathogenic variants, we can determine which of these regions are most relevant to particular diseases and identify regions poorly suited to evolutionary analysis, like gain-of-function hotspots.

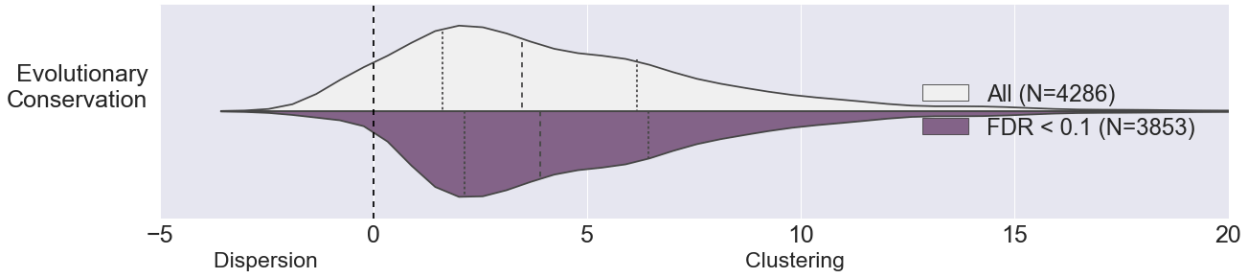

Figure 6: Distribution of protein z-scores for the weighted univariate analysis of evolutionary conservation as measured by Jenson-Shannon divergence. Evolutionary conservation is significantly clustered in 88% of protein structures.

Pathogenic missense variants are spatially clustered within protein structures

Pathogenic missense variants derive their pathogenicity from the disruption of protein structure and function. We have demonstrated that functional residues commonly form clusters in protein structure, thus we expect pathogenic variants to form clusters at functional sites relevant to

specific diseases. We analyzed 449 protein structures containing three or more pathogenic variants from ClinVar and identified a global trend towards clustering (Figure 5), with significant clustering of pathogenic variants in 125 proteins (22%) and significant dispersion in one protein. To determine if these clusters were characteristic of pathogenic variation – and not spatial constraints on missense variation – we performed a bivariate analysis of pathogenic variants relative to putatively neutral missense variants from ExAC in 440 proteins containing three or more variants from each set. We identified 112 proteins (25%) in which ClinVar pathogenic variants were significantly more clustered than ExAC missense variants, 98 of which were also significant in the univariate analysis (Figure 6). This union represents proteins in which the clustering of ClinVar pathogenic variants is statistically significant and independent from the general patterns observed for missense variation, suggesting that pathogenic variants have focal effects within a protein structure.

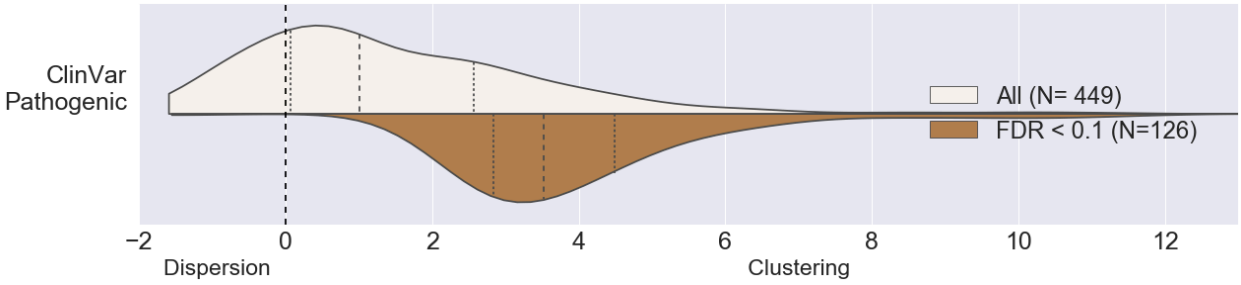

Figure 7: Distribution of z-scores for the un-weighted univariate analysis of ClinVar pathogenic missense variants. Pathogenic variants demonstrate a strong trend towards spatial clustering, with significant clustering identified in 126 proteins (28%).

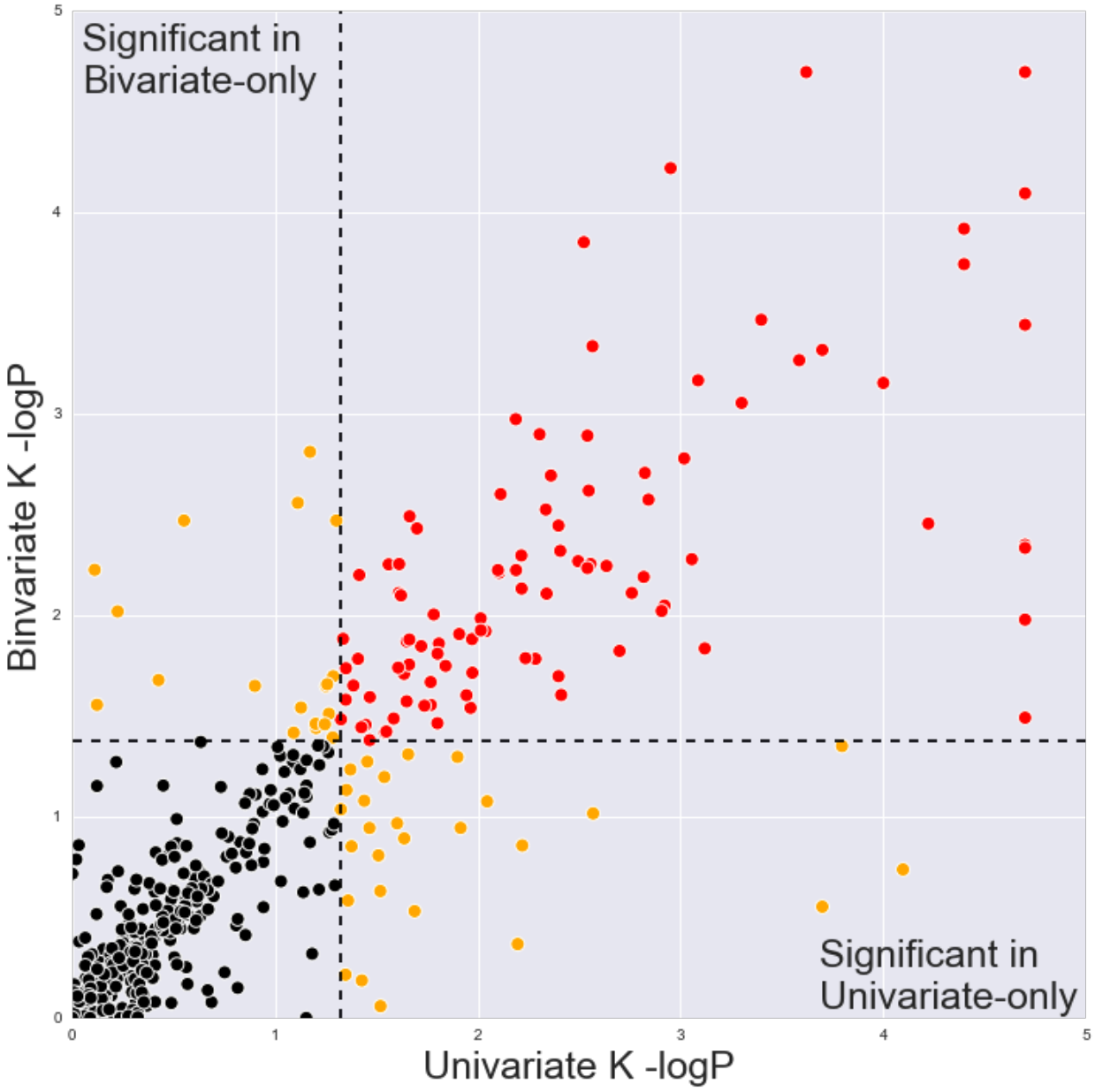

Figure 8: Comparison of the univariate and bivariate p-values for ClinVar pathogenic variation. The dashed lines mark the p-value threshold from each analysis where q<0.1. Proteins plotted in orange were significantly clustered in only one analysis. Proteins plotted in red had significant clustering of ClinVar pathogenic variants that exceeded what was observed for ExAC missense variants.

Dominant missense variants form smaller clusters than recessive variants

Having demonstrated pathogenic variants trend towards spatial clustering, we next evaluated whether the mode of inheritance for pathogenic variants influences their spatial constraint. Missense variants causing protein loss-of-function (LoF) may disrupt numerous critical elements of a protein structure, but the opportunity for gain-of-function (GoF) is likely limited to a small subset of regions with functional potential. Previous work by Turner *et al*.[23] investigated these spatial patterns in protein sequence using autosomal dominant (AD, typically gain-of-function) and autosomal recessive (AR, typically loss-of-function) missense variants from the Human Gene Mutation Database[36] (HGMD). Turner *et al.* demonstrated a significant global trend for dominant variants to be more clustered than recessive, which in turn were more clustered than neutral variants from the 1000 Genomes Project, with dominant variants in 9 proteins and recessive variants in 5 proteins significantly more clustered than neutral variants (FDR<5%).

The functional impact of gain- and loss-of-function missense variants is derived from their effect on protein structure. Thus, the spatial distributions derived from these effects are perhaps more accurately evaluated within that context. Using the HGMD dataset curated by Turner *et al*., we performed two bivariate analyses evaluating dominant and recessive missense variants relative to ExAC missense variants (Figure 7). We identified 27 (of 69, 39%) and 16 (of 47, 34%) structures in which dominant and recessive variants (respectively) were significantly more clustered than variants from ExAC (FDR<10%). Additionally, we found that univariate scores for both dominant and recessive variants were significantly higher (more clustered) than ExAC variants (AD: $p=3.53x10^{-30}$, AR: $p=6.97x10^{-20}$ Mann Whitney U), but found no significant difference between dominant and recessive variants (p=0.274). However, within proteins with significantly clustered variation, dominant variants ($N_{AD}=35$) formed significantly smaller

clusters (median peak significance: 10Å) than recessive variants (N_{AR}=16) (median peak significance: 13.5Å) (p=0.014 Mann Whitney U). These findings support previous conclusions that both gain- and loss-of-function variants are more clustered than neutral variants. The smaller clusters formed by dominant variants additionally support the hypothesis that gain-of-function mutations are localized to specific sites with functional potential, while loss-of-function mutations more generally disrupt regions of functional importance.

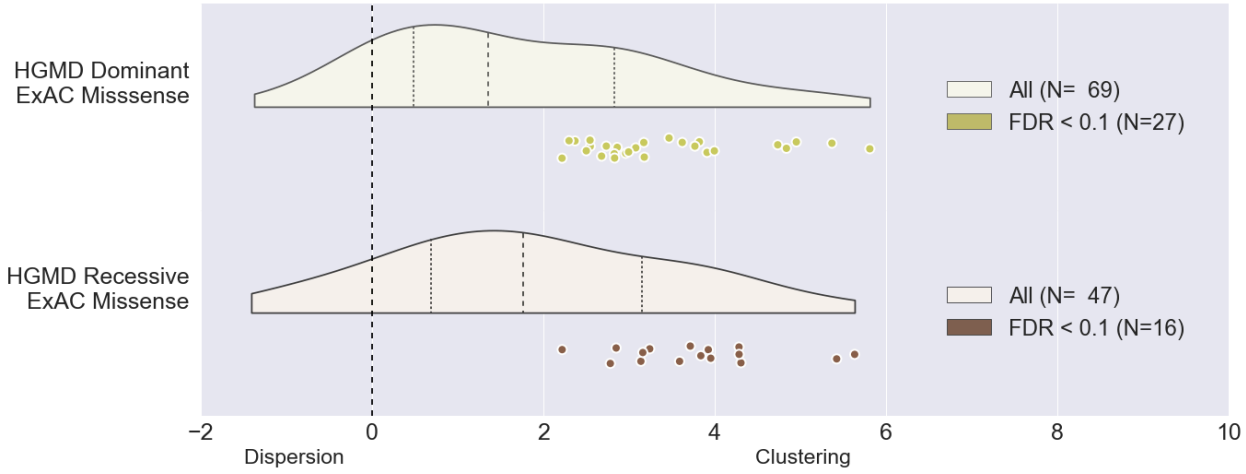

Figure 9: Autosomal dominant and recessive missense variants from the Human Gene Mutation Database (HGMD) are both spatially more clustered than ExAC missense variants in protein structure, consistent with ClinVar pathogenic variation. No significant difference in the strength of clustering was identified between the two groups, but dominant mutations did on average form smaller clusters (AD=11Å, AR=14Å).

Proximity to clustered pathogenic variants is predictive of pathogenicity

The identification of pathogenic variant clusters in protein structures may lead to better understandings of disease etiology and improvements in pathogenicity prediction for variants of unknown significance. To estimate the predictive potential of spatial information, we defined a simple metric that ranks amino acids by their relative proximity to pathogenic and neutral variation and measures predictive performance using leave-one-out cross validation (Figure 8).

Applying this approach to all 442 proteins from the bivariate analysis of ClinVar pathogenic and ExAC missense variants did not accurately classify pathogenic and neutral variants (median ROC AUC=0.55), but performance on the subset of 98 proteins with significant univariate and bivariate clustering of pathogenic variants was significantly improved (median ROC AUC=0.73, $p=2.46x10^{-17}$ Mann Whitney U) and comparable to SIFT, PolyPhen2, and evolutionary conservation (ANOVA p=0.128) (Figure 9). To determine if protein fold or conformation influenced the performance of spatial proximity, we stratified our analysis by CATH domain. We observed no significant difference in performance between protein structures containing the various CATH domains (Figure 10). These results suggest that proximity to pathogenic clusters, not individual pathogenic variants, is predictive of variant pathogenicity.

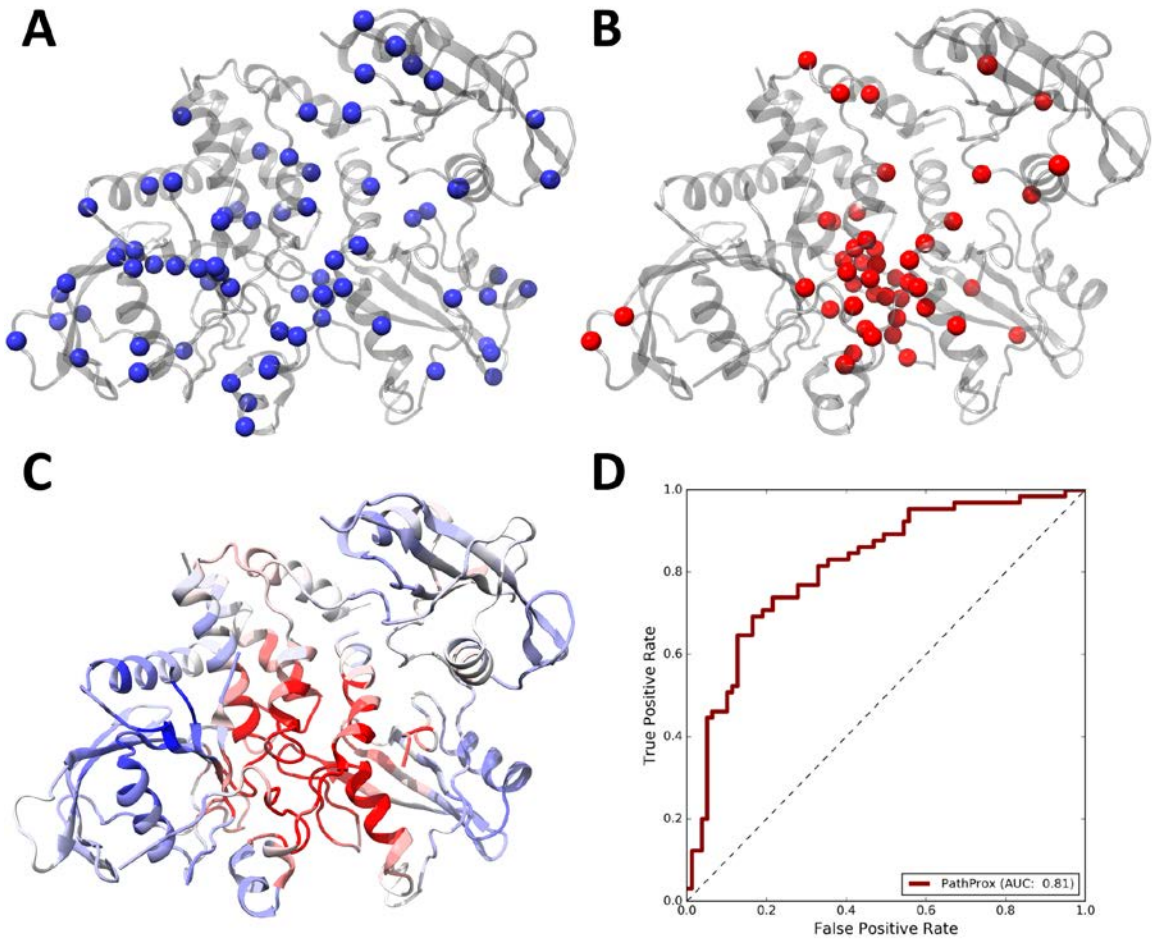

Figure 10: The average proximity of each residue to missense variants in the (A) neutral and (B) pathogenic datasets is measured using the NeighborWeight function. Residues are then scored by (C) the difference in their pathogenic and neutral proximity scores. Prediction performance is then quantified using the area under the (D) receiver-operating characteristic curve (ROC AUC).

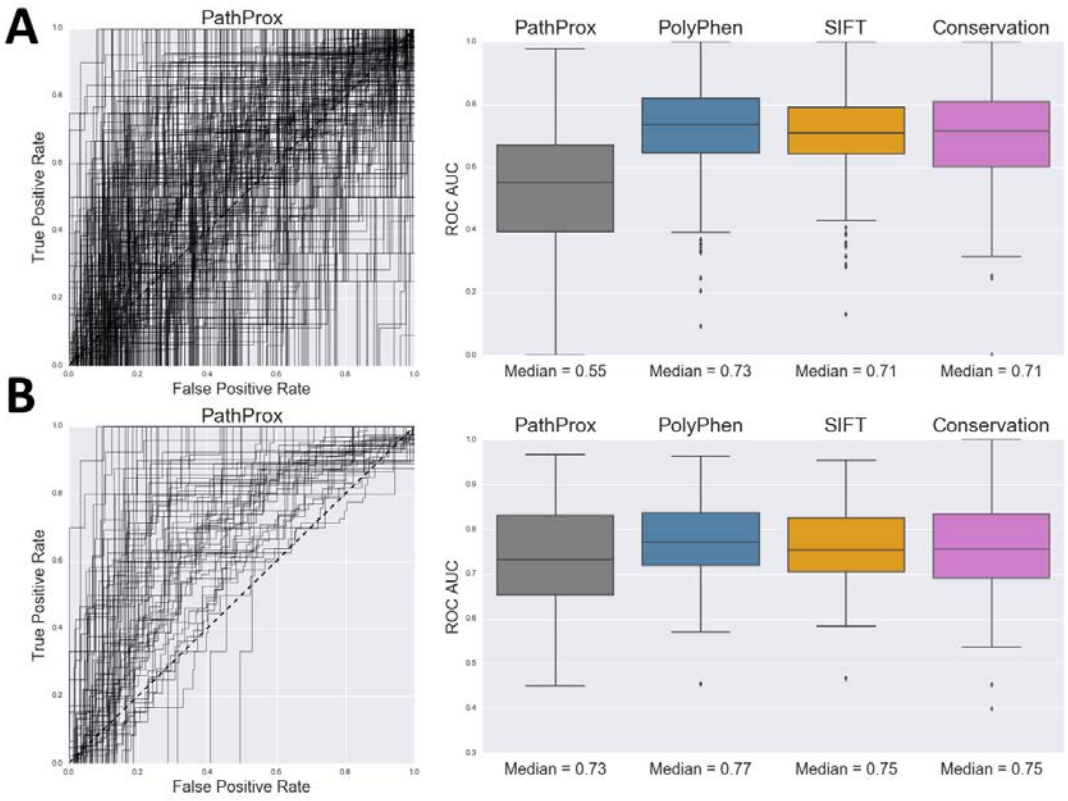

Figure 11: Receiver-operating characteristic (ROC) curves for spatial prediction performance. (A) Predictive performance over all proteins is poor, but (B) for proteins in which ClinVar pathogenic variants are significantly more clustered than ExAC missense variants, spatial prediction performance is comparable (ANOVA p=0.46) to general pathogenicity prediction algorithms.

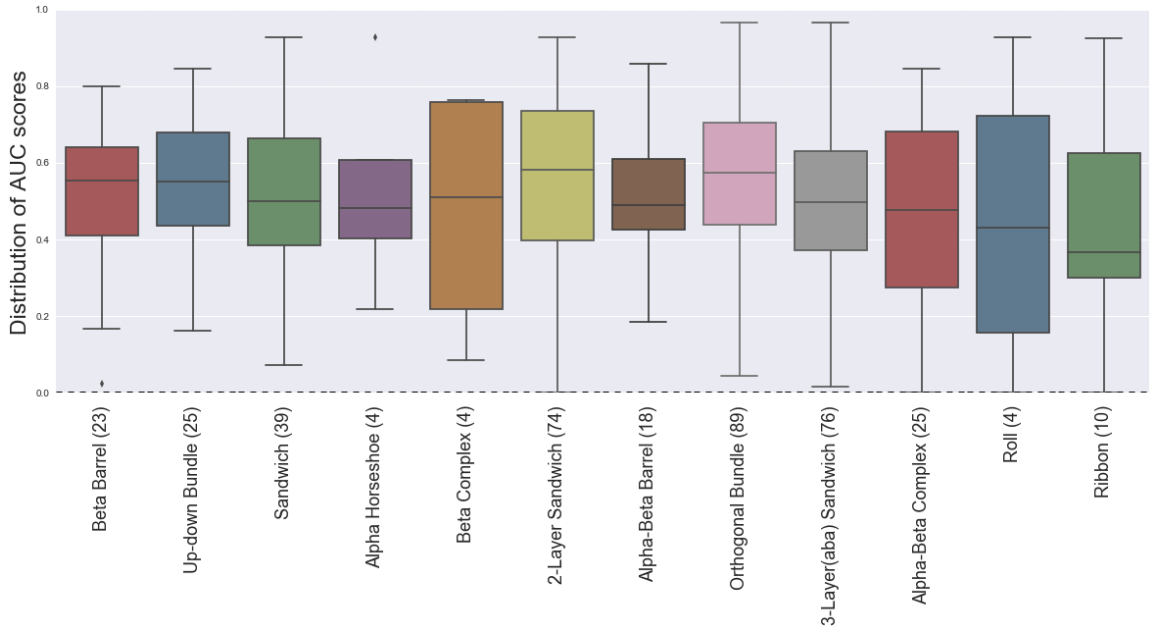

Figure 12: Spatial prediction performance stratified by CATH domain. There were no significant differences in prediction performance between CATH domains (ANOVA p=0.28).

CHAPTER IV

DISCUSSION

The near-random spatial distributions of synonymous variation, in contrast with the trend towards spatial dispersion observed for missense variants, suggests that non-random spatial patterns of missense variation are derived from the functional effects (or lack thereof) of amino acid substitutions. Pathogenic variants are often significantly clustered in protein structure and are typically more clustered than neutral missense variants, with dominant variants forming smaller, more localized clusters than recessive variants. In proteins where pathogenic variants are significantly clustered and significantly more clustered than neutral variants, simple spatial predictors perform as well as general predictors of variant pathogenicity.

The spatial clustering of evolutionarily conserved amino acids in protein structures has been shown previously in only a small number of proteins. Our analysis evaluated all proteins with solved structures and confirmed the trend on a large scale. This property of conserved residues supports the hypothesis that deleterious variants, which are presumed to disrupt or enhance protein function, will also form spatial clusters.

Within most proteins, the difference between the spatial distributions of synonymous and missense variants distributions is subtle. However, comprehensive analysis of all proteins with solved structures reveals a significant shift towards missense variant dispersion, while synonymous variants vary little from spatial randomness. This shift towards dispersion and the significantly non-random spatial distributions of missense variants in 51 proteins suggest that the neutral and deleterious effects of amino acid substitutions influence the spatial constraint we

observe for missense variation. It follows that we can infer from these distributions which regions of a protein are variant-intolerant. For proteins with significantly dispersed missense variation, we infer that variation in the core of the protein, likely affecting protein stability, is not well tolerated. Similarly, proteins with significant clustering of missense variants are variant-intolerant in general, but contain isolated regions where mutations can be introduced without deleterious effects.

While conservation analysis provides a hypothesis about which regions of a protein are evolutionarily important and vulnerable to disruption, clusters of pathogenic variants identify regions with relevance to human disease. Pathogenic gain-of-function variant clustering may also identify regions with functional potential not captured by evolutionary analysis. We find that pathogenic missense variants were significantly clustered in 28% of the analyzed proteins. Comparison with the neutral missense background further improves our ability to identify pathogenic clustering that exceeds neutral expectation, while simultaneously filtering proteins in which the observed pathogenic clustering is attributable to general patterns of missense variation. In total, 99 proteins were found to contain significant clustering of pathogenic missense variants that significantly exceeded what was observed for neutral missense variants.

Separately analyzing dominant and recessive variants demonstrates that this phenomenon is not limited to gain- or loss-of-function. The tendency for dominant variants to form smaller clusters supports previous findings and suggests that gain-of-function potential is limited to a small number of residues, while loss-of-function variants affect larger regions of existing functional importance. These differences may assist in the classification of variants of unknown significance by setting an expectation for the distance between disease-causing variants in a

protein. Candidate variants within a plausible distance of a pathogenic cluster may be prioritized over more distant variants.

In the presence of pathogenic clustering, spatial proximity has the potential to enhance variant pathogenicity prediction. Clusters of pathogenic variation indicate regions of a protein structure that are functional, intolerant to variation, and contributory to disease. When attempting to classify variants of unknown significance, variants in close proximity to such a cluster are more likely to perturb functional and thus have a higher likelihood of causing a similar, deleterious effect. The limitation of this approach to variants with significantly clustered pathogenic variation may be due to small samples of known, pathogenic variants, but is likely exacerbated by a lack of specificity when relying exclusively on spatial information. The inclusion of evolutionary and biochemical features is essential to properly assessing the potential effect of an amino acid substitution.

To comprehensively investigate spatial patterns of genetic variation, we have summarized the multi-distance results for each protein by calculating the area between expected and observed K values. This approach captures robust signals identifiable over many distances. Other summarization approaches may capture different spatial patterns or reveal signals that are significant only at very specific distance scales. Ultimately, the utility of our multi-distance approach is the fine-grained analysis of individual protein structures, such that the details of these patterns are apparent. For an individual protein-of-interest, we recommend a review of the multi-distance results to identify the patterns most relevant to a specific hypothesis. Further, we have analyzed only a representative subset of proteins with solved structures in the Protein Data Bank, but selecting a structure in a relevant conformation is critical for the meaningful interpretation of spatial patterns within that protein.

We have demonstrated in this study that evolutionarily conserved residues are almost universally clustered in protein structures, that missense variants are generally dispersed, that pathogenic variants are generally (and often significantly) clustered, and that differences between these distributions have the potential to predict variant pathogenicity. The methods we have proposed identify patterns in spatial data that are analogous to the input parameters required by many density-based and hierarchical clustering algorithms. Using data-driven parameterization of these algorithms, we can accurately define spatial boundaries for evolutionarily conserved, disease-associated, or variant-intolerant regions of protein structures. We have also started developing novel variant-aggregation and association methods that incorporate the spatial relationships identified in this study. In combination with the wealth of phenotypic data in electronic health records, which are becoming increasingly accessible for research, these approaches have the potential to associate specific protein regions directly with clinical outcomes. This thesis makes a significant advance in our understanding the spatial distribution of missense variants in protein structures that will ultimately improve our understanding of the link between protein structure, function, and human disease.

REFERENCES

1. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins*. 2011;79(3):830-838. doi:10.1002/prot.22921.Role.

2. Baugh EH, Simmons-Edler R, Mueller CL, et al. Robust Classification of Protein Variation Using Structural Modeling and Large-Scale Data Integration. *Preprint*. 2015;XX(Xx):1-6. doi:10.1093/nar/gkn000.

3. Goldenberg O, Erez E, Nimrod G, Ben-Tal N. The ConSurf-DB: Pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res*. 2009;37(SUPPL. 1):323-327. doi:10.1093/nar/gkn822.

4. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput Biol*. 2009;5(12). doi:10.1371/journal.pcbi.1000585.

5. Madabushi S, Yao H, Marsh M, et al. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J Mol Biol*. 2002;316(1):139-154. doi:10.1006/jmbi.2001.5327.

6. Mihalek I, Reš I, Lichtarge O. A Family of Evolution-Entropy Hybrid Methods for Ranking Protein Residues by Importance. *J Mol Biol*. 2004;336(5):1265-1282. doi:10.1016/j.jmb.2003.12.078.

7. Schueler-furman O, Baker D. Conserved Residue Clustering and Protein Structure Prediction. 2003;235(July 2002):225-235.

8. Mirny L a, Shakhnovich EI. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J Mol Biol*. 1999;291(1):177-196. doi:10.1006/jmbi.1999.2911.

9. Landgraf R, Xenarios I, Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J Mol Biol*. 2001;307(5):1487-1502. doi:10.1006/jmbi.2001.4540.

10. Panchenko AR, Kondrashov FA, Bryant S. Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci*. 2004;13:1-9. doi:10.1110/ps.03465504.and.

11. Manning JR, Jefferson ER, Barton GJ. The contrasting properties of conservation and correlated phylogeny in protein functional residue prediction. *BMC Bioinformatics*. 2008;9:51. doi:10.1186/1471-2105-9-51.

12. Del Sol Mesa A, Pazos F, Valencia A. Automatic methods for predicting functionally important residues. *J Mol Biol*. 2003;326(4):1289-1302. doi:10.1016/S0022-2836(02)01451-1.

13. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. *Nat Struct Biol*. 1995;2(2):171-178. doi:10.1038/nsb0295-171.

14. Thusberg J, Olatubosun A, Vihinen M. Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat*. 2011;32(4):358-368. doi:10.1002/humu.21445.

15. Li B, Krishnan VG, Mort ME, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*. 2009;25(21):2744-2750. doi:10.1093/bioinformatics/btp528.

16. Bao L, Zhou M, Cui Y. nsSNPAnalyzer: Identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res*. 2005;33(SUPPL. 2):480-482. doi:10.1093/nar/gki372.

17. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7(4):248-249. doi:10.1038/nmeth0410-248.

18. Bromberg Y, Rost B. SNAP: Predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res*. 2007;35(11):3823-3835. doi:10.1093/nar/gkm238.

19. Thomas PD, Campbell MJ, Kejariwal A, et al. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res*. 2003;13(9):2129-2141. doi:10.1101/gr.772403.

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

21. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*. 2002;30(17):3894-3900. doi:10.1093/nar/gkf493.

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

23. Turner TN, Douville C, Kim D, et al. Proteins linked to autosomal dominant and autosomal recessive disorders harbor characteristic rare missense mutation distribution patterns. *Hum Mol Genet*. 2015;24(21):5995-6002. doi:10.1093/hmg/ddv309.

24. Abecasis GR, Auton A, Brooks LD, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65. doi:10.1038/nature11632.

25. Meyer MJ, Lapcevic R, Romero AE, et al. Mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. *Hum Mutat*. 2016:n/a-n/a. doi:10.1002/humu.22963.

26. Kamburov A, Lawrence MS, Polak P, et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. 2015:1-10. doi:10.1073/pnas.1516373112.

27. Collin Tokheim, Rohit Bhattacharya, Noushin Niknafs, Derek M Gygax, Rick Kim M, Ryan, David Masica RK. Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res*. 2016. doi:10.1158/0008-5472.

28. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*. 2013;29(18):2238-2244. doi:10.1093/bioinformatics/btt395.

29. Araya CL, Cenik C, Reuter JA, et al. Systematic identification of significantly mutated regions reveals a rich landscape of functional molecular alterations across cancer genomes. *Nat Genet*. 2016:20875. doi:10.1101/020875.

30. Stehr H, Jang S-HJ, Duarte JM, et al. The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. *Mol Cancer*. 2011;10:54. doi:10.1186/1476-4598-10-54.

31. Niu B, Scott AD, Sengupta S, et al. Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat Genet*. 2016;(June). doi:10.1038/ng.3586.

32. Guo Y, Wei X, Das J, et al. Dissecting disease inheritance modes in a three-dimensional protein network challenges the "guilt-by-association" principle. *Am J Hum Genet*. 2013;93(1):78-89. doi:10.1016/j.ajhg.2013.05.022.

33. Forbes SA, Beare D, Gunasekaran P, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res*. 2015;43(Database issue):D805-11.

34. Chang K, Creighton CJ, Davis C, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;45(10):1113-1120. doi:10.1038/ng.2764.

35. Lek M. Analysis of protein-coding genetic variation in 60,706 humans. 2015:1-26. doi:http://dx.doi.org/10.1101/030338.

36. Stenson PD, Ball E V., Mort M, et al. Human Gene Mutation Database (HGMD): 2003 Update. *Hum Mutat*. 2003;21(6):577-581. doi:10.1002/humu.10212.

37. Berman HM. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235-242. doi:10.1093/nar/28.1.235.

38. Pieper U, Webb BM, Barkan DT, et al. ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res*. 2011;39(Database issue):D465-74. doi:10.1093/nar/gkq1091.

39. Landrum MJ, Lee JM, Benson M, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*. 2016;44(D1):D862-D868. doi:10.1093/nar/gkv1222.

40. Feyfant E, Sali A, Fiser A. Modeling mutations in protein structures. *Protein Sci*. 2007;16(9):2030-2041. doi:10.1110/ps.072855507.

41. Velankar S, Dana JM, Jacobsen J, et al. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res*. 2013;41(Database issue):D483-9. doi:10.1093/nar/gks1258.

42. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*. 2010;26(16):2069-2070. doi:10.1093/bioinformatics/btq330.

43. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-842. doi:10.1093/bioinformatics/btq033.

44. Dixon PM. Ripley's K function. *Encycl Environmetrics*. 2002;3(December):1796-1803. doi:10.1002/9780470057339.var046.

45. Diggle PJ, Chetwynd a G. Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*. 1991;47(3):1155-1163. doi:http://www.jstor.org/stable/2532668.

46. Gaines KF, Bryan a L, Dixon PM. The Effects of Drought on Foraging Habitat Selection of Breeding Wood Storks in Coastal Georgia. *Waterbirds*. 2000;23:64-73. doi:10.2307/4641111.

47. Cunningham F, Amode MR, Barrell D, et al. Ensembl 2015. *Nucleic Acids Res*. 2014;43(D1):D662-669. doi:10.1093/nar/gku1010.

48. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res*. 2014;43(D1):D204-212. doi:10.1093/nar/gku989.

49. Cock PJA, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25(11):1422-1423. doi:10.1093/bioinformatics/btp163.

50. Capra JA, Singh M. Predicting functionally important residues from sequence

conservation. *Bioinformatics*. 2007;23(15):1875-1882. doi:10.1093/bioinformatics/btm270.

51. Touw WG, Baakman C, Black J, et al. A series of PDB-related databanks for everyday needs. *Nucleic Acids Res*. 2015;43(D1):D364-D368. doi:10.1093/nar/gku1028.

52. Durham E, Dorr B, Woetzel N, Staritzbichler R, Meiler J. Solvent accessible surface area approximations for rapid and accurate protein structure prediction. *J Mol Model*. 2009;15(9):1093-1108. doi:10.1007/s00894-009-0454-9.

53. Moutsianas L, Agarwala V, Fuchsberger C, et al. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet*. 2015;11(4):e1005165. doi:10.1371/journal.pgen.1005165.

54. TG R, HA S, MA R, et al. A Protein Domain and Family Based Approach to Rare Variant Association Analysis. 2016:e0153803. doi:10.1371/journal.pone.0153803.

55. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89(1):82-93. doi:10.1016/j.ajhg.2011.05.029.

56. de Beer T a P, Laskowski R a, Parks SL, Sipos B, Goldman N, Thornton JM. Amino Acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. *PLoS Comput Biol*. 2013;9(12):e1003382. doi:10.1371/journal.pcbi.1003382.