Constraint on Rare Protein-Coding Variation: Pathogenicity Prediction and Phenotypic Discovery

By

Robert Michael Sivley

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Informatics

January 31, 2018

Nashville, Tennessee

Approved:

John A. Capra, PhD

William S. Bush, PhD

Jens Meiler, PhD

Jonathan A. Kropski, MD

Antonis Rokas, PhD

DEDICATION

I never planned to pursue genetics or computational biology, so I would like to dedicate this dissertation to everyone who prepared me for, guided me towards, and supported me throughout this endeavor.

To my parents, you instilled in me a curiosity about the world and the confidence to pursue my goals and aspirations without fear of the unknown.

To Lance, you pinned a request for a research assistant on a crowded poster board that changed my career trajectory. When you went out of your way to follow up with my graduate school plans, I told you that I was looking for a research assistantship to help pay for my tuition; you introduced me to my future PhD advisor.

To Tricia, you took a risk and hired a Computer Science Master's student when you could have hired someone with more genetics experience. You introduced me to genetics, tolerated my ignorance in the early days, and turned a pay-the-bills research assistantship into a career in scientific research.

To Will, you showed me that the work that I was doing was publishable, and that a career in science could take me places during graduate school that I never expected to see in my lifetime.

To Tony, you looked past my insecurities and tab-delimited file explanation and saw the potential for a scientist.

To Alex, you gave me a reason to stay in Nashville. You gave me a reason to travel the world. You gave me the support I needed to pursue a graduate degree in a completely new field. I may not have attempted this without you; I *could not* have accomplished this without you.

When I first began working with the Center for Human Genetics, I had at best a high school understanding of biology. This dissertation is the product of a series of incredibly fortunate events and several incredible people. I owe all of this to you. Thank you.

ACKNOWLEDGMENTS

This dissertation would not be possible without the training and mentorship provided by my advisors, Drs. Will Bush and Tony Capra. When I was looking for programmer positions, Will and Tony encouraged me to join the graduate program and pursue a PhD. They have guided me along the path from programmer to scientist. They have taught me about biology and genetics, about machine learning, about scientific rigor and experimental design, and how to write a scientific manuscript. I also owe a great deal to Drs. Jens Meiler and Jonathan Sheehan, who have acted as my teachers, collaborators, and co-authors in all things structural biology, a topic of critical importance to the work presented here and in my previous work.

I would also like to thank all members of the Bush and Capra labs for their tremendous support and friendship. Alex Fish, Corinne Simonti, and I were founding members of the Capra lab, and we leave it behind with complete confidence that its current membership will carry the lab to greatness. In the same vein, I'd like to thank the human genetics graduate students in the Vanderbilt Genetics Institute for accepting a Biomedical Informatics graduate student into the fold.

Finally, I would like to acknowledge the financial support I have received from the ocular genomics training grant (5T32EY021453-05), the Vanderbilt Ingram Cancer Center SPORE, and the Undiagnosed Disease Network.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

Introduction

1.1 Motivation

The ultimate goal of studying human genetics is to understand genome function and its effects on the human phenotype. One avenue by which to study the genome is through patterns of genetic variation within modern populations, which provide insight into functional and evolutionary constraints on different loci. For example, a lack of common genetic variation in a locus is often indicative of functional constraint against variation, suggesting that sequence changes negatively influence reproductive fitness[1]. Similar to how co-segregation of a genetic marker with disease is evidence of disease association, the absence of variation across thousands of healthy individuals is evidence that variation in the locus is associated with poor health. By this logic, we can begin to interrogate the human genome for regions potentially contributing to human disease using data derived from the general population.

The first systematic examinations of fully sequenced human genomes established consistently stronger constraint (i.e., less genetic variation) in protein-coding regions compared to non-coding sequences[2–5]; exons harbor approximately half the level of genetic variation as introns and non-coding flanking sequences. Furthermore, early candidate gene sequencing studies identified lower rates of non-synonymous variation than synonymous variation within protein-coding regions[6], highlighting the increased constraint on protein-altering mutations. Patterns of constraint observed in the general population are driven by selective pressures on that population, and may serve as markers to indicate which parts of the genome are most important for human health. By quantifying these patterns with measures of selective constraint, we can potentially use this information to interpret the phenotypic effects of rare and novel genetic mutations.[7,8]. Building on exome-sequencing data from over one hundred thousand individuals, we are now able to expand beyond targeted sequencing and whole-genome genotyping of disease-specific pedigrees and cohorts and quantify constraint genome-wide, across the frequency spectrum, and for individuals representing many continental and regional ancestries. In this dissertation, we explore three avenues by which constraint on protein-coding variation can be used to better understand human biology and elucidate the genetic drivers of disease. We first integrate human genetics with protein structural biology to identify proteins with evidence for constraint on where germline variation is and is not tolerated; we then use this information to predict whether new variants will contribute to disease. Next, we investigate somatic mutations from human tumors to identify proteins with evidence of constraint on the location of these potentially cancer-driving variants. Finally, we investigate genes that, across tens of thousands of individuals, appear significant less tolerant to loss-of-function mutations; using genetic data from hospital patients, we explore the phenotypic impact of these genes and to better understand their importance to human health.

1.2 Background

The basic dogma of protein structural biology is that genes encode mRNA, mRNA is translated into a protein sequence, and the protein sequence folds into a three-dimensional conformation by which it performs its function, either independently, in conjunction with other proteins, or with other copies of the same protein. A genetic variant that changes a protein's sequence can impact its overall three-dimensional conformation, disrupt the morphology of functional domains, or alter local biochemical environments necessary for interactions with other molecules. When we study the phenotypic impact of protein-coding genetic variation, we are ultimately studying the phenotypic impact of conformational and biochemical changes to a protein structure.

For many proteins, we have experimentally-determined models of their three-dimensional structures. These models are only a snapshot, and represent one of potentially many biologically relevant conformations, but they provide invaluable information about protein function. The Protein Data Bank[9] serves as the central repository for experimentally derived protein structural information, and currently includes data for nearly 25% of human proteins. For those proteins without structural information, computational techniques exist to predict their three-dimensional conformations using the known structures of homologous proteins. Large-scale endeavors like ModBase[10] aim to predict protein structure for the entire proteome, filling the information gap between the proteome and structome and increasing partial coverage of the human proteome to approximately 75%. While computational approaches produce structures of lesser quality and are typically insufficient for precise energetics calculations, they provide a good interim source of structural information suitable for many other tasks.

By integrating human genetics with structural biology, we can study protein-coding genetic variation within its functional context: protein structure. The workflow for mapping a protein-coding variant into a protein structure is conceptually straightforward: a genetic variant alters a protein-coding nucleotide, which alters the translated amino acid, that amino acid is mapped to a specific position in the protein sequence, and that position is then mapped to a specific coordinate in a relevant protein structure. In practice, this process is complicated by the integration of multiple, independently maintained databases and resources that are often in disagreement with one another. Furthermore, performing the task of mapping genetic variants into protein structures becomes computationally demanding at scale. Thus, the task of integrating entire genetic datasets—many derived from whole-exome sequencing and containing hundreds of thousands of protein-coding variants—required the creation of a novel resource and database. In previous work, I developed PDBMap, a high-throughput pipeline for the integration of genetics and structural biology, capable of mapping entire genetic datasets into all solved and predicted protein structures while identifying and correcting for disagreements between the component datasets. With this resource, we are able to investigate patterns of constraint on protein-coding genetic variation directly within protein structure.

Also in previous work, we developed a spatial statistic—based on Ripley's K[11]—for measuring the degree of clustering or dispersion of missense variation

within protein structure, which is discussed in detail in Chapter 3. Briefly, the approach identifies significantly non-random spatial patterns of genetic variants in protein structure. It can recognize localized clustering of variants as well as larger spatial patterns, like variant depletion within a structural domain. Using this measure, we can identify proteins in which the spatial distribution of genetic variation appears to be under significant constraint. We have already made several interesting discoveries with this approach; in our analysis of germline disease-causing variation from ClinVar[12] and population-derived, putatively neutral variation from the Genome Aggregation Database[7] (gnomAD), we found that germline disease-causing variants are often significantly clustered within protein structure, while variants from the general population are often spatially dispersed, typically occupying the protein surface where variation is better tolerated. Using a simple metric—spatial proximity to known disease-causing variants—we discovered that non-random spatial patterns have the potential to predict variant pathogenicity. These findings, summarized in Chapter 1, motivated us to explore whether constraint on variation in protein structure could classify variants of unknown significance.

Predicting the pathogenicity of variants of unknown significance

Variants in many genes are known to contribute to heritable diseases, and this information is useful for clinical diagnosis. For example, prenatal screening of the gene Phenylalanine-4-hydroxylase (PAH) can identify mutations known to cause phenylketonuria (PKU), a severe pediatric disease in which individuals cannot metabolize phenylalanine. Without treatment, phenylketonuria leads to a build up of phenylalanine in the brain, and the development of permanent intellectual disability. While many mutations in PAH are known to cause PKU, the challenge comes when sequencing identifies a variant that has not been shown to cause PKU; novel and rare variants identified in disease-associated genes are usually classified as variants of unknown significance (VUS). For a VUS in PAH, a clinician may suggest a diet low in phenylalanine and monitor phenylalanine levels over time to determine if the patient has PKU. The handling of VUS becomes more challenging for genes like the breast cancer-associated genes BRCA1 and BRCA2, where incorrect classification of a VUS could mean that a low-risk patient undergoes a major, but unnecessary surgery.

As clinical sequencing of the human exome becomes more affordable, interpreting the clinical impact of VUS is becoming a major challenge for physicians. Sequencing technologies provide information on all genetic variants, including rare and novel mutations; the clinical relevance of these mutations is often unknown, even when they occur within genes with known disease associations. Improvements in our ability to assess variant pathogenicity in the clinic has the potential to improve diagnosis of the underlying cause of disease in patients and guide personalized treatment, Thus, the ability to differentiate between variants that do and do not cause disease is of paramount importance to genetic testing, precision medicine and the future of clinical care.

A number of algorithms provide predictions for missense pathogenicity by analyzing patterns of evolutionary conservation and/or biochemical characteristics

of amino-acid substitutions. SIFT[13] is one of the earliest predictors, and also one of the most popular. Using a combination of multiple sequence alignment and the BLOSUM62 amino acid substitution matrix, SIFT calculates the probability that an amino acid will be tolerated within a given protein sequence. A close rival to SIFT, PolyPhen2[14] also uses multiple sequence alignment, but incorporates sequence information like protein domain boundaries and structural information like solvent accessibility, and derives its predictions using a naïve Bayesian classifier trained on either HumDiv (a dataset of variants causing Mendelian disease and putatively neutral sequence differences between humans and mammalian homologs) or HumVar (a dataset of human disease-causing variants and putatively neutral common variants from human populations). While popular, disagreement between these algorithms is frequent; in one report, the correlation between SIFT and PolyPhen2 scores was only 0.4[16]. Newer pathogenicity prediction methods have also attempted to incorporate protein structure more directly; VIPUR[17] adds computational structural biology to sequence- and structure-derived features to estimate changes in thermodynamic stability caused by a mutation, however it only slightly improves prediction performance at the expense of runtime. Several other pathogenicity prediction methods are discussed and compared in a recent review, which found the accuracy of all methods considered to range between 0.60 and 0.82[15]. While these methods make frequent use of evolutionary conservation and in some cases protein structural information, none consider spatial patterns of variation within human protein structures, which provide information on functional constraints, and may improve prediction performance.

As stated earlier, we found that the spatial distributions of disease-causing and non-disease-causing variants are distinct from one another and have the potential to predict the pathogenicity of uncharacterized variants. To evaluate the practical application of this approach to the classification of variants of unknown significance in a clinical setting, we focused our efforts on variation in the regulator of telomere elongation helicase 1 (*RTEL1*). *RTEL1* is responsible for telomere maintenance, and the dysregulation of *RTEL1* is associated with familial idiopathic pneumonia (FIP). Short telomeres are frequently observed in patients with FIP, but the specific mechanisms by which variation in *RTEL1* causes FIP is still poorly understood, thus predicting the effects of VUS is quite challenging; some variants are tolerated while others lead to dramatic alterations in protein structure, trafficking/localization, or function[15]. Classical genetic approaches, including linkage analysis, may also be confounded for telomere-related genes by the inheritance of short telomeres (and thus increased disease risk) without inheritance of the causal allele. Assigning pathogenicity to VUS has important implications for genetic testing and family counseling.

We present a novel approach that utilizes the distribution of disease-causing variants in the protein structure of RTEL1 for pathogenicity prediction. The analysis uses the spatial distribution of variants of known effect to classify variants of unknown significance.

Constraint on the spatial distribution of somatic mutations in cancer

We have so far exclusively discussed selective constraint in the context of germline variation. Cancer genomics presents an interesting challenge in the measurement of constraint because many of our basic assumptions are either incorrect or reversed. For example, germline variants are subject to selective constraint within every tissue, throughout the entire lifespan of an organism, and across generations. Thus, germline variants rising to high frequency in the general population must not negatively impact reproductive fitness. In contrast, somatic mutations in a tumor context arise in a highly dysregulated environment, impact only the tissue in which they occur, and are subject to very different selective pressures; variants that increase the fitness of tumor cells, rather than the whole organism, are positively selected. Somatic mutations are inherently a mixture of two mutation types. Driver mutations cause or promote the progression of cancer, and are important for determining personalized treatment. The identification of driver mutations is complicated by the presence of passenger mutations, which arise as a result of the dysregulated tumor context and do not necessarily contribute to the development of the cancer. Ultimately, protein-coding somatic driver mutations are not dissimilar to protein-coding germline pathogenic variation. Both variant types cause phenotypic effects that negatively impact organismal fitness, and each often do so by altering the structure and/or function of a protein. By comparing the mutational landscape between individuals with cancer, we can identify regions of proteins more often affected by somatic mutations than expected at random.

Indeed, clusters of highly recurrent somatic mutations in both protein sequence and structure have been observed in several thoroughly studied cancer genes and are believed to be a hallmark of driver mutations. It follows that the identification of this spatial signature in other genes may identify previously unknown cancer genes and isolate regions of proteins most relevant to tumorigenesis.

Detecting cancer genes and driver mutations with somatic mutation clustering

Recognizing the importance of large reference datasets for cancer, two projects have compiled data on tens of thousands of somatic mutations. The Catalogue of Somatic Mutations in Cancer[18] (COSMIC) is a submission-based system for somatic mutations observed in individuals with cancer. This resource provides an unparalleled amount of mutational data, but submission-based ascertainment, targeted sequencing of known cancer genes, and selective reporting of mutations has potentially biased the data available for analysis. For tasks that require an unbiased population reference, data from The Cancer Genome Atlas[19] (TCGA) is more appropriate. All data from TCGA is derived from whole-exome sequencing studies of individuals with specific cancer types. This approach has yielded less overall data, but provides a somatic reference dataset that can be used for general cancer research, or for comparison between different types of cancer.

Several recent studies in cancer have searched for spatial clusters of somatic mutations. The first systematic assessment of somatic mutation clustering across many protein structures was performed by Stehr *et al.*[20], who calculated the sum of

inverse pairwise distances between somatic mutations from COSMIC. This approach was not designed to identify specific mutation clusters, but rather to determine whether somatic mutations were significantly clustered in the protein structures of known oncogenes and tumor suppressors (as catalogued by COSMIC at the time of publication). Although the analysis included only 24 proteins, they conclude that somatic mutations in oncogenes, but not tumor suppressors, were significantly more clustered (were more nearby one another) than common, population-derived variants from the 1000 Genomes Project[21]. Their findings suggested that the spatial clustering was primarily a characteristic of gain-of-function somatic mutations. This conclusion was challenged by Kamburov *et al*.[22] three years later. Using whole-exome sequencing data from TCGA and an improved measure of spatial clustering that considered mutation recurrence, they quantified the degree of somatic mutation clustering in a comprehensive set of 4,062 proteins. Their approach also used a transformation of the Euclidean distance between mutations to up-weight mutations at biologically meaningful distances (~5Å) and down-weight mutations at greater distances (~15Å). Counter to the previous findings by Stehr *et al.*, this approach identified significant clustering in both oncogenes and tumor suppressor genes, suggesting that spatial analysis of somatic mutations may be of broad relevance in cancer genetics. To further improve the clinical utility of somatic cluster analysis, Meyer *et al.*[23], Tokheim *et al.*[24], and Niu *et al.*[25] each developed clustering algorithms to identify specific clusters of somatic mutations. In contrast to the approaches described above, which quantify clustering of variants within a protein structure as a function of their proximity to one another, a formal clustering algorithm has the benefit of detecting mutation hotspots and structural regions of tumorigenic importance with specificity; rather than stating that variants are clustered, this approach identifies the specific mutations that form those clusters. Cluster-based approaches have many similarities; Meyer *et al.* and Niu *et al.* both employ hierarchical clustering algorithms, Tokheim *et al.* and Niu *et al.* both identify hotspot mutations before defining their clusters. These and other methods are discussed in a recent review of methods for detecting cancer driver mutations[26].

Despite the subtle differences in the underlying methodology and dataset selection of these studies, all have identified somatic mutation clustering in either protein sequence[27,28] or structure[20,22–25]. However, the genes identified have been inconsistent across studies. We expect this discordance is due primarily to differences in three factors: (1) methods for identifying (and defining) clusters of somatic mutations, (2) protein structural dataset selection, and (3) somatic mutation dataset selection. Using our previously developed framework for measuring spatial distributions (described in detail in Chapter 3), we reanalyzed previously considered genetic and structural datasets with a consistent methodology to quantify the prevalence of somatic clustering in protein structures and to identify a high-confidence, high-coverage set of proteins with significant spatial constraint on somatic mutations in cancer.

Quantifying selective constraint at the gene level

Ultimately, the analysis of protein-coding variation within its structural context will provide the highest resolution information about the functional effects of disease-causing variants. However, we can also use measures of selective

constraint in the absence of known disease associations to identify regions of functional importance. Much in the way that evolutionary conservation is informative as to what parts of the genome have been important across evolutionary timescales, patterns of genetic variation in modern populations provide information about where variation is and is not tolerated within the human genome.

One approach to measuring constraint at the gene-level is the Residual Variance Intolerance Score[29] (RVIS), which uses data from the Exome Sequencing Project[4] (ESP). RVIS works by regressing the amount of observed common missense and protein-truncating variants on the total number of variants (including rare and synonymous) to identify genes with large studentized residuals; large residuals are interpreted as either an increase in or relaxation of selective pressures for that gene. This approach performs well in identifying genes associated with Mendelian traits, in particular those associated with dominant-negative and haploinsufficient phenotypes.

A slightly more recent method, the probability of loss-of-function (LoF)-intolerance[7] (pLI), uses whole-exome sequencing data from the Exome Aggregation Consortium[30] (ExAC) and focuses specifically on protein-truncating variants (PTVs), including nonsense, splice acceptor, and splice donor variants; i.e. variants expected to disrupt translation of the gene transcript, resulting in an incomplete and putatively non-functional protein. To calculate the pLI for a gene, the number of missing protein-truncating variants (PTVs) is first estimated. The probability of observing all possible protein-truncating mutations is calculated by simulating all possible single nucleotide mutations and evaluating whether that mutation would be protein-truncating. This count is then adjusted by the estimated mutation rate of the gene, as derived from the regional divergence between humans and macaques[8]. Finally, the expected number of PTVs is estimated from number the number of possible PTVs, the estimated mutation rate, the sample size, and sequencing depth (observing a variant is inherently less likely in poorly sequenced regions). A score is defined that quantifies the amount of missing PTVs per gene, defined as the one minus the observed number of PTVs divided by the expected number of PTVs, such that scores approaching one are indicative of loss-of-function intolerance. Using this score, genes are then assigned using the expectation-maximization algorithm to one of three groups: genes tolerant of LoF variation, genes for which heterozygous PTVs are tolerated, and haploinsufficient genes (heterozygous PTVs are not tolerated). The ratio of observed to expected PTVs in the latter two categories is estimated from known recessive and haploinsufficient disease genes. Finally, the probability of LoF-intolerance (pLI) is defined as the likelihood that a gene is belongs in the haploinsufficient category; genes with a pLI$\geq$0.9 are predicted to be LoF-intolerant[7].

Genes predicted to be LoF-intolerant are broadly and highly expressed, are depleted for expression quantitative-trait loci (eQTL), enriched for core biological pathways, have many physical interaction partners, and include most haploinsufficient disease genes; in short, LoFi genes appear to very important. We would thus expect genetic variants affecting these genes to increase risk for human disease. However, despite abundant evidence for functional importance and selective constraint, 72% of the 3,230 LoFi genes are not currently associated with

any human disease phenotype. A recent set of meta-analyses of whole-exome sequencing studies aimed to quantify the impact of PTVs in LoF-intolerant genes for 13 quantitative traits and 10 human diseases[31]. Genome-wide burden of PTVs in LoF-intolerant genes was significantly associated with increased risk for several psychiatric disorders, including bipolar disorder, autism, schizophrenia, intellectual disability, and attention deficit hyperactivity disorder. Genome-wide burden analysis of predicted pathogenic missense variants in LoF-intolerant genes identified similar associations, suggesting that deleterious missense variants and protein-truncating variants may contribute to similar phenotypic outcomes. Other complex diseases, like type II diabetes and inflammatory bowel disease, showed no association with the burden of PTVs in LoF-intolerant genes. This study benefited from the availability of whole-exome sequencing data for over 100,000 individuals, but not all phenotypes were available for all samples, limiting the power of each individual association analysis. Furthermore, while whole-exome sequencing data provides an excellent source for low-frequency and *de novo* PTVs, focusing these analyses exclusively on PTVs in genes known to be depleted for PTVs greatly reduces the amount of available data, lowering the likelihood of identifying significant associations and effectively eliminating the possibility of gene-level analysis, which this study did not perform. Finally, phenotypic information for only a small number of quantitative traits and diseases means that this study did not fully interrogate the phenotypic spectrum.

One interpretation for the extreme constraint observed for LoF-intolerant genes is that any disruption of their function leads to deleterious phenotypes detrimental for survival or reproduction[7]. This implies not only that different PTVs can cause the deleterious effect, but also that these effects could be caused by mutations of other types, like missense variants. Thus, a model of allelic heterogeneity is most appropriate when testing for significant associations with disease. In the study described above, two aggregate association strategies were used: burden analysis and sequence-kernel association testing (SKAT). Burden tests assume a logistic relationship between the number of variants observed in an individual and the likelihood of developing a disease or trait. This approach is statistically powerful when the assumption is true and all variants are associated with the phenotype and affect risk in the same direction: either increasing or decreasing, collectively. This assumption can be relaxed using SKAT to accommodate variants impacting the phenotype in opposing directions: some increasing and some decreasing, as well as variants with neutral effects[32]. In practice, the truth of this underlying assumption is often unknown, so an optimized method (SKAT-O) was developed to balance the contribution of each approach and maximize statistical power[33]. Aggregate association statistics are invaluable when many low-frequency causal variants (as opposed to a single causal common variant) are hypothesized to affect risk for a disease; e.g. cystic fibrosis can be caused by many variants within *CFTR* [MIM: 602421]. This hypothesis is especially relevant for diseases caused by protein loss-of-function (rather than gain-of-function) because of the many ways of disrupting protein function, making it an ideal choice for interrogating the phenotypic impact of variation in loss-of-function intolerant genes.

Although the extreme level of constraint on LoF variation suggests the

importance of LoF-intolerant genes, and while disruption of these genes is expected to have severe phenotypic consequences, the pLI metric does not supply any hypotheses regarding which phenotypes to expect; it is very likely that the most common phenotype for LoF variants in LoF-intolerant genes is embryonic lethality. However, we hypothesize that variants causing less severe disruptions to protein function are associated with human disease, and that these diseases can provide new insights into why the complete disruption of LoF-intolerant genes is incompatible with life. To interrogate a spectrum of possible phenotypes for a gene, we perform a Phenome Wide Association Study[34,35] (PheWAS). This approach is typically enabled by a clinical biobank linked with an electronic medical record (EMR), from which we can derive both genetic and broad phenotypic information from clinical samples. Standard phenotyping algorithms for PheWAS aggregate groups of related ICD9 billing codes into PheWAS codes that reflect clinical phenotypes. The assignment of case, control, and exclusion status for each sample is determined by the presence or absence of the relevant billing codes. Because billing codes are not diagnoses, and may often be assigned during routine course of treatment before a final diagnosis is made, phenotyping algorithms typically require multiple occurrences of the same code for a sample to be classified as a case. Once the case/control status of each sample is determined for each phenotype, a PheWAS is conducted as a series of independent association tests, similar to a genome-wide association study (GWAS). When multiple loci are included, a PheWAS is conducted for each locus. Like GWAS, PheWAS involves thousands to millions of association tests, so correction for multiple testing is critical to avoid spurious findings.

Although access to the electronic-medical record provides an invaluable catalogue of phenotypic information, the clinical cohorts used for PheWAS introduce several caveats and considerations. The medical record is not a closed-world system, so it is possible that some controls used for any given analysis are not true controls; no mention of a disease does not preclude the possibility that an individual was diagnosed elsewhere or will develop the disease later in life. Because clinical populations are often curated from a single hospital, there is also an increased risk that study findings will not generalize beyond the hospital or region in which the study was performed. Similarly, a regionally curated clinical population is likely to include many families, which must be considered for any statistical analysis assuming un-related individuals. For these reasons, it is advisable to view the results of a PheWAS as data-driven hypotheses requiring additional statistical and biological support. Despite these caveats, EMR-linked biobanks provide the most comprehensive phenotypic information available for analysis, and have been successful in identifying novel phenotypic associations with uncharacterized genes and variants of predicted functional importance[35–38].

Using dense whole-exome genotyping in an EMR-linked biobank, we use gene-level aggregate association tests to interrogate all protein-coding variants in predicted LoF-intolerant genes for association with clinically derived phenotypes.

1.3 Chapters

In chapter 2, we hypothesize that disease-causing missense variants in the protein structure of RTEL1 are spatially clustered in functionally important regions that, when disrupted, contribute to the development of pulmonary fibrosis. We evaluate this hypothesis using spatial analytics that measure the degree of clustering or dispersion of variants within protein structures, and use spatial information to classify missense VUS identified through Sanger sequencing of *RTEL1* in families with familial idiopathic pneumonia.

In chapter 3, we hypothesize that somatic cancer-driver mutations exhibit spatial patterns similar to what has been previously observed for germline disease-causing variants, due to similar contributions to human disease. Because driver and passenger mutation data is limited, we analyze all somatic mutations for spatial clustering, using recurrence to enrich our dataset for potential driver mutations. We explore the spatial distribution of somatic mutations from two data databases of cancer genetics, in the context of solved and predicted protein structures. Finally, we compare and contrast the results of our analysis with previous comprehensive analyses of somatic mutation clustering protein structure.

In chapter 4, we shift our focus to genes without existing associations with human disease, but for which whole-exome sequencing of healthy individuals suggests biological and phenotypic importance. Using a collection of genes predicted to be loss-of-function intolerant, we use gene-level aggregate analysis of rare variation to perform a phenome-wide association study (PheWAS) using clinical phenotypes derived from the electronic medical record. We compare and contrast the phenotypic associations of loss-of-function tolerant and intolerant genes, present novel gene-phenotype associations, and demonstrate existing mouse model support for several of the significant associations.

In chapter 5, we discuss the implications and limitations of this dissertation, and propose future work, including the adaptation of a new measure of selective constraint for protein structure, and the replication of our PheWAS results using whole-exome sequencing data from a large EMR-linked biobank.

Chapter 2

Classifying variants of unknown significance in RTEL1 using spatial constraint

The content of this chapter is adapted from a submitted manuscript: Sivley, R.M., Sheehan, J., Kropski, J., Cogan, J., Blackwell, T.S., Phillips, J.A., Bush, W.S., Meiler, J., Capra, J.A., Three-dimensional spatial analysis of missense variants in RTEL1 identifies pathogenic variants in patients with Familial Interstitial Pneumonia. In revision.

2.1 Introduction

For many genes, associations with disease are limited to a handful of genetic variants that collectively fail to explain the heritability of the disease[39]. Often, the presence of multiple disease-causing variants within a single gene suggests that the missing heritability may be explained by additional, currently uncharacterized genetic variants. However, large-scale population sequencing has made it abundantly clear that not every variant in a disease-associated gene causes that disease[7,21]. By analyzing the patterns of constraint on disease-causing and putatively neutral variants, in particular patterns of spatial constraint in protein structure, we can identify regions of proteins in which mutations are more or less likely to cause disease and help to elucidate the underlying mechanism of disease. In previous work, we demonstrated the predictive potential of this approach; we now aim to explore its practical application to variants of unknown significance identified in a clinical setting.

The use of next-generation sequencing to study families with pulmonary diseases has led to the identification of novel genes and mechanisms associated with the inherited forms of pulmonary arterial hypertension[40–44] and pulmonary fibrosis[45–47]. Genetic variation in telomere-related genes is the predominant cause of pulmonary disease (when genetic etiology is known). Even when the genetic cause is unknown, such as with idiopathic pulmonary fibrosis, telomere shortening in peripheral blood mononuclear cells[48–50] and type II alveolar epithelial cells[45,50] is commonly observed in patients and families. The mechanism through which telomere dysfunction leads to lung fibrosis is not clear, but may involve premature senescence of progenitor cells in the distal lung[51–53]. Among families with pulmonary fibrosis (familial interstitial pneumonia, FIP), whole exome sequencing (WES) studies have identified that variation in a few genes is responsible for disease risk. The most commonly mutated genes in FIP patients are *TERT* (10–15% of cases)[54,55], *RTEL1,* and *PARN* (3–4% of cases each)[45,46]. Most FIP mutations identified to date are very rare or novel. Rare variation presents challenges when using genetic information in clinical practice, since most newly identified variants in FIP-associated genes are considered variants of unknown significance (VUS).

We screened FIP families from our registry for rare variants in *RTEL1* and identified 13 rare missense VUS. We hypothesized that pathogenic *RTEL1* variants likely affect critical functions and/or protein interactions and thus would co-

localize in three-dimensional space. To test this hypothesis, we used homology modeling to predict the tertiary structure of RTEL1 and identified a spatial cluster of variants with known disease-association in RTEL1's helicase domains. We then developed an algorithm to classify missense VUS based on their spatial proximity to known pathogenic and neutral variants with the expectation that VUS near the pathogenic cluster are more likely contribute to disease. The approach outperformed two common pathogenicity prediction methods in cross-validation and predicted the pathogenicity of disease-segregating VUS with high accuracy. Our study supports the likely pathogenicity of novel FIP-associated rare variants, generates a new homology model of RTEL1's 3D structure, supports quantitative spatial analysis in protein structure as a powerful approach to classify VUS in *RTEL1,* and suggests this technique may have broad applicability to other genes and genetic diseases.

2.2 Methods

Subjects and Samples

We trained our spatial proximity prediction algorithm using putatively neutral *RTEL1* missense variants from the 1000 Genomes Project[21] that were not otherwise associated with disease and pathogenic missense variants causing severe pediatric, autosomal recessive Hoyeraal-Hreidarsson syndrome collected from previous literature[56–62]. We evaluated the performance of our prediction algorithm using rare missense variants of unknown significance from patients with Familial Interstitial Pneumonia (FIP). Subjects were identified from the Familial Interstitial Pneumonia (FIP)/Familial Pulmonary Fibrosis (FPF) registries at Vanderbilt University, the University of Colorado, and National Jewish Hospital[45]. FIP was defined by the presence of Idiopathic Interstitial Pneumonia (IIP) in two or more family members, including IPF in at least one individual. Phenotypes of subjects selected for sequencing were ascertained using ATS/ERS criteria for IIP[63]. The affected status of deceased individuals was determined by review of available medical records, autopsy material, or by death certificates. DNA was isolated from blood and/or paraffin-embedded lung tissue using a PureGene Kit (Gentra Systems, Minneapolis, MN). Rare missense variants (MAF < 0.001) in *RTEL1* were curated from whole-exome sequencing data as previously reported [45] (n=189 families) or targeted modified Sanger sequencing of *RTEL1* (n=184 families) (Figure 1). Co-segregation and telomere length measurements were performed as previously described[45]. VUS co-segregation with disease and short telomeres was considered evidence for pathogenicity and represent true-positives in our analysis.

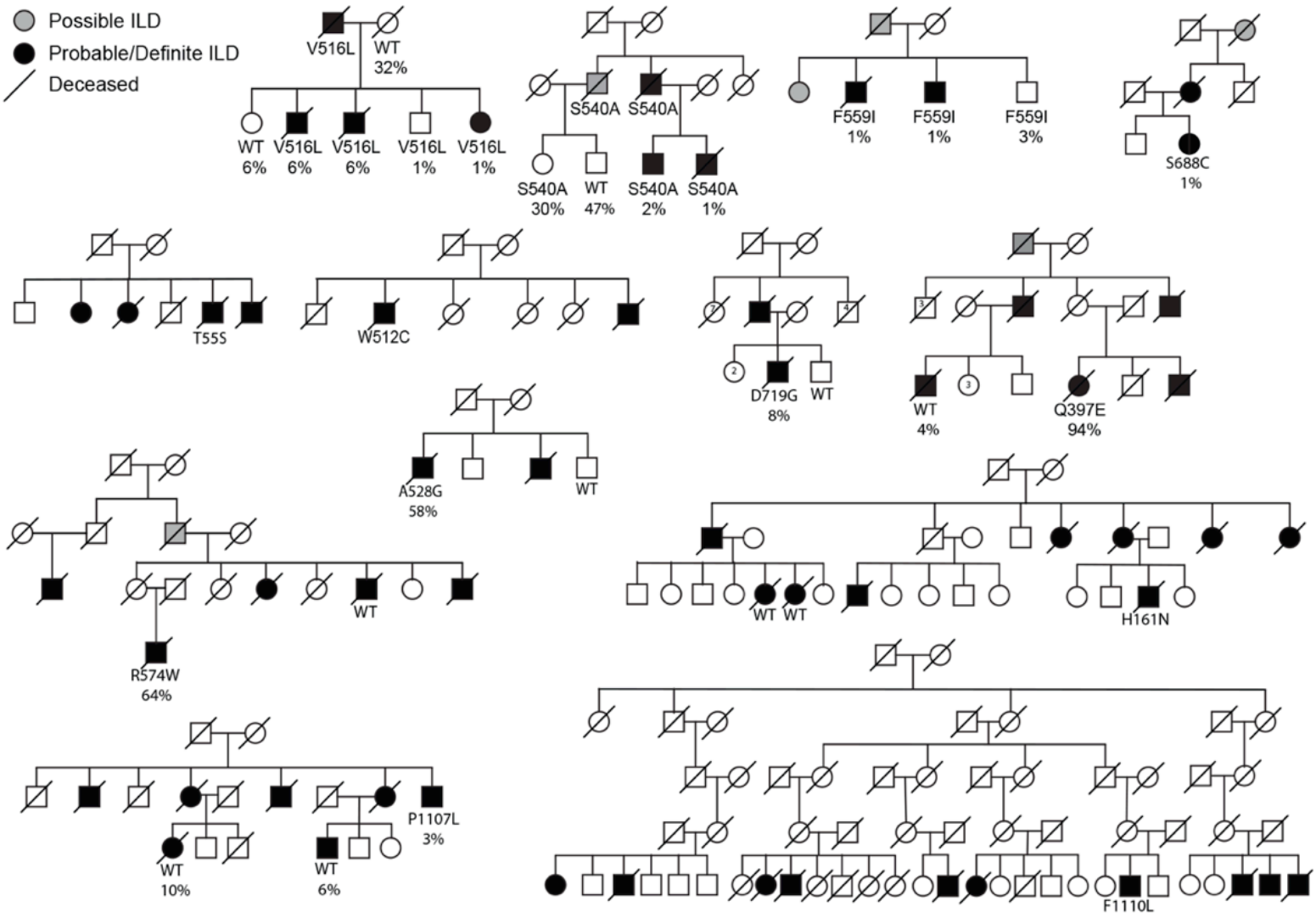

Figure 1: Familial Idiopathic Pneumonia (FIP) pedigrees. Genotyping of 373 FIP patients identified 13 missense variants of unknown significance (VUS) in RTEL1. Analysis of pedigrees of FIP patients demonstrated that seven VUS segregate with disease. Telomere percentages are provided below each mutation when available.

Protein Structural Analysis

We quantified the spatial proximity of each VUS to each known pathogenic and neutral variant using the NeighborWeight transformation of the 3D Euclidean distance between the centroid of each amino acid side chain[64],

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

where $d_{x,y}$ is the distance between VUS x and variant y from set Y (pathogenic or neutral) and the bounds give upper and lower bounds in angstroms. This transformation up-weights the contribution of nearby variants and down-weights

distant variants that are less likely to have similar functional effects (Figure 2). To capture neighboring residues with the potential for direct interaction, the lower bound was set to 8 Å. The upper bound was set to 24 Å to capture variants potentially impacting the same functional domain or element. We then calculated the proximity P of each VUS x to variants in dataset Y using the weighted-average of transformed distances,

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

To classify VUS, we calculated the difference in the pathogenic and neutral proximity scores,

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

such that candidate VUS in closer proximity to pathogenic variation than neutral variation receive positives scores. We refer to ΔP as the pathogenic proximity score.

Figure 2: Neighbor Weight as a function. The Neighbor Weight function transforms distance using a lower bound of 8Å and an upper bound of 24Å to up-weight nearby variants and downweights distant variants.

We evaluated the predictive power of the pathogenic proximity score using leave-one-out cross-validation on the known pathogenic and neutral variants [65]; each variant was predicted to be pathogenic or neutral by its proximity to all other variants. We quantified the performance of each prediction method using the area under the receiver operating characteristic curve (ROC AUC). The ROC curve plots true positive rate, the proportion of true positives (pathogenic variants) predicted to be positive, versus false positive rate, the proportion of true negatives (neutral variants) predicted to be positive, as a function of prediction rank. The ROC AUC is equivalent to the probability that a randomly selected positive is ranked higher than a randomly selected negative; thus, perfect separation of positives and negatives produces a ROC AUC of 1.0 and random ordering produces a ROC AUC of 0.5. We compared the performance of the pathogenic proximity score with other pathogenicity prediction methods, including ConSurf evolutionary conservation scores [66], SIFT [13], and PolyPhen2 [14]. A brief description of each approach is provided in the Supplemental Methods.

Pathogenicity prediction methods

ConSurf [66] calculates the relative evolutionary conservation of each amino acid within a protein sequence, with scores ranging from –1.5 (most conserved) to 1.5 (least conserved). Variants were ranked by increasing ConSurf score; amino acid substitutions at the most conserved residues were predicted to be deleterious, and at the least conserved residues to be neutral. SIFT [13] and PolyPhen2 [14] are machine learning classifiers designed to predict the impact of amino acid substitutions using a combination of sequence- and/or structure-derived features. SIFT uses multiple sequence alignments of closely-related homologs to calculate the probabilities of amino acid substitutions. Substitutions with low likelihoods are predicted to be deleterious while those with moderate or high likelihoods are predicted to be benign. PolyPhen2 provides posterior probability estimates of pathogenicity ranging from 0 (benign) to 1 (damaging). SIFT and PolyPhen2 scores were calculated by the Ensembl Variant Effect Predictor [67].

2.3 Results

Constructing a structural model of RTEL1

The protein structure for RTEL1 has not yet been experimentally determined, so we constructed a computationally derived homology model. To begin, we applied nine computational modeling algorithms to the protein sequence: GeneSilico [68], HHpred [69], I-TASSER [70], M4T [71], Pcons5 [72], Phyre2 [73], RaptorX [74], Robetta [75], and SWISS-MODEL [76]. RaptorX produced the highest-coverage model, which consisted of two well-folded domains spanning residues 1-769 and 881-1151. This model was based on seven PDB structures: 4a15 [77], 3crv [78], 2fi7 [79], 2gm7 [80], 4pjq [81], 2vrw [82], 4a64 [83]. To improve quality, the model was relaxed using Rosetta version 2015.19 [84], and then subjected to 1000 rounds of loop_modeling [85] using perturb_kic_with_fragments.

Known pathogenic missense variants in RTEL1 cluster in 3D structure

To analyze the 3D distribution of disease-associated RVs in *RTEL1*, we mapped known pathogenic and neutral variants onto the sequence and structure of RTEL1 (Figure 1). Because the relative orientation of the N- and C-terminal models (residues 1-769 and 881-1151) is unknown, we analyzed variants in these models separately. In the N-terminal model, we observed spatial clustering of pathogenic variants in helicase domain II (Figure 3a) and near the structural interface of helicase domains I and II (Figure 1b). This tendency was not observed among neutral variants, which were distributed throughout the protein structure. The distinct spatial distributions of pathogenic and neutral variation suggest that clustering is characteristic of pathogenic variation in *RTEL1* and that disease-causing missense RVs in *RTEL1* disrupt similar protein functions. In the C-terminal model there were relatively few candidate VUS relative to the N-terminal model, leading to poor performance (Figure 4). We focused the remainder of our analyses on the N-terminal model.

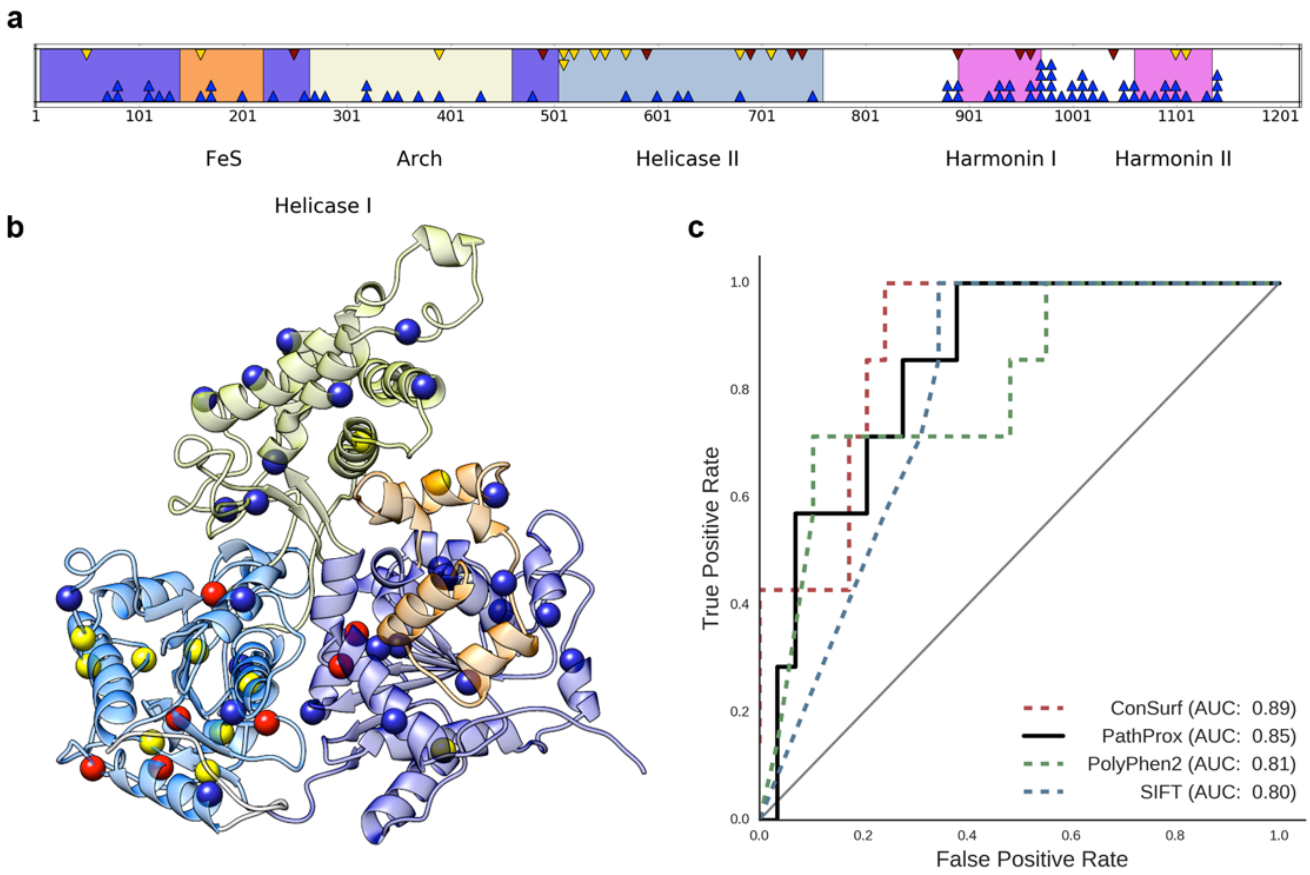

Figure 3: Identification and classification of novel pathogenic FIP variants in RTEL1. (a) The locations of known pathogenic (red), putatively neutral 1000 Genomes (blue), and FIP VUS (yellow) missense variants are plotted in the context of the RTEL1 protein sequence and known domains. (b) The locations of pathogenic, putatively neutral, and candidate variants in the RTEL1 N-terminal structural model. (c) Leave-one-out cross validation of the pathogenic proximity score applied to characterized *RTEL1* variants yielded an improved area under the ROC curve (AUC) relative to PolyPhen2 and SIFT, but was outperformed by evolutionary conservation scores. These results demonstrate that considering the 3D spatial distribution of known pathogenic and neutral variants can identify pathogenic hotspots and assist in the classification of VUS.

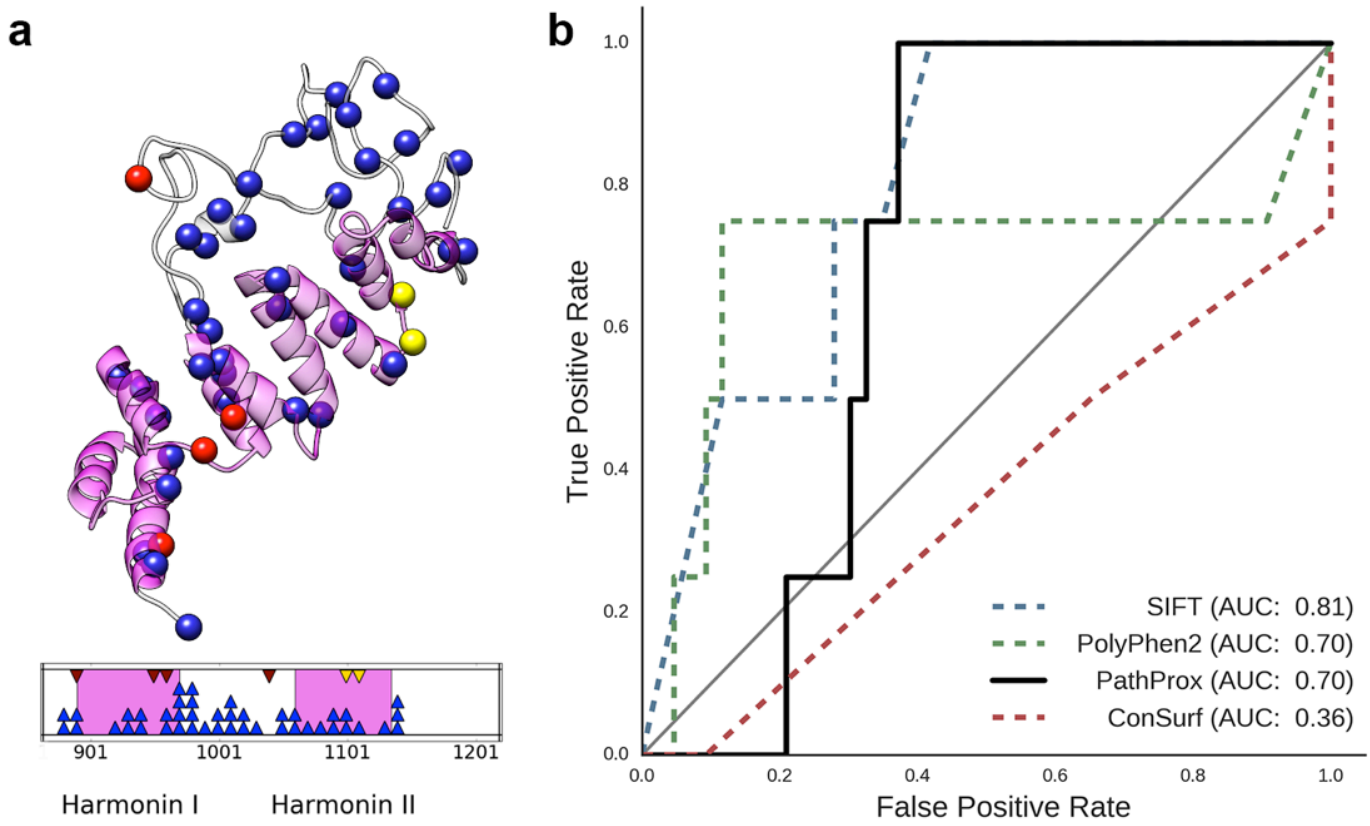

Figure 4: Reduced performance of spatial proximity in the variant-sparse C-terminal model of RTEL1. (a) The locations of known ClinVar pathogenic (red), putatively neutral 1000 Genomes (blue), and new candidate FIP (yellow) missense variants in the RTEL1 C-terminal structural model. (b) Receiver operating characteristic (ROC) curves for variants in the C-terminal model of RTEL1. ConSurf did not provide conservation scores for most residues in the C-terminal model. Only four pathogenic variants were present in the C-terminal model and predictive performance was notably worse than in the larger N-terminal model.

Spatial proximity analysis accurately classifies pathogenic and neutral RTEL1 variants

Based on the observed differences between neutral and pathogenic variant distributions, we hypothesized that candidate VUS could be classified by their relative spatial proximity to known pathogenic and neutral variants. To evaluate this, we used leave-one-out cross-validation to calculate pathogenic proximity scores (ΔP) for each known pathogenic and neutral variant in the N-terminal model of RTEL1 and then plotted ROC and PR curves to measure how accurately the proximity score predicts pathogenicity. Classifying variants by their pathogenic proximity score performed well (Figure 1c); the approach yielded a ROC AUC of 0.85.

To estimate the sensitivity of the proximity-based prediction method to the number of known pathogenic variants, we recomputed pathogenic proximity scores using all possible subsets of pathogenic variants, and then calculated the ROC and PR AUC for each subset (Figure 5). As expected, performance increases as the number of known pathogenic mutations considered increases; the mean ROC AUC is 0.62 when only two pathogenic variants are known and 0.82 when six variants are considered. This suggests that performance will increase as more

pathogenic variants are identified. However, we caution that the number of known pathogenic variants required will likely vary substantially based on the structure and function of the protein of interest.

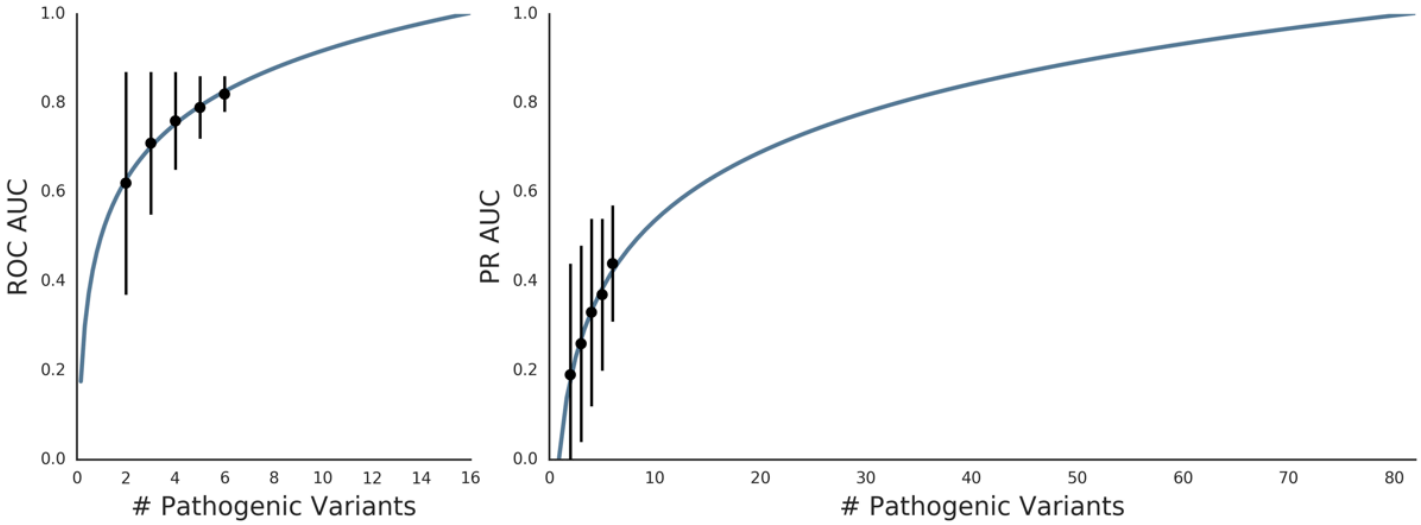

Figure 5: Estimation of the sensitivity of pathogenic-proximity-based prediction to the number of known pathogenic variants. ROC AUC (a) and PR AUC (b) were computed based on all subsets of the seven known pathogenic variants in RTEL1's N-terminal domain. The black dots represent the mean performance across all subsets of each size (subsets of 2 to 6 variants) and the bars represent the standard deviation. The blue curve was fit to the log-AUC using linear regression.

We then compared the performance of our pathogenic proximity score to a representative set of current methods for *in silico* pathogenicity prediction: ConSurf evolutionary conservation [66], SIFT [13], PolyPhen2 [14] (Figure 1c). The pathogenic proximity score outperformed PolyPhen2 (ROC AUC=0.81) and SIFT (ROC AUC=0.80); evolutionary conservation had the best performance (ROC AUC=0.89). The competitive ROC AUC with current methods and the relatively strong performance obtained with small numbers of known pathogenic variants demonstrates the predictive potential of spatial statistics, which are not currently used for variant pathogenicity prediction.

The pathogenic proximity score identifies nearly all disease-segregating VUS as pathogenic

Given the predictive potential of the pathogenic proximity score, we applied our methodology to the 13 missense VUS identified from our FIP registry; six that segregate with disease, five that do not segregate with disease, and two for which segregation data was unavailable. The pathogenic proximity score classified eight VUS as deleterious (Table 1), including five VUS (V516L, S540A, F559I, S688C, D719G) that co-segregated with disease and were found in subjects with short telomeres in peripheral blood mononuclear cells, a biomarker of reduced RTEL1 activity [48–50] (Figure 1). Two false positives (A528E, R574W) did not co-segregate with disease or were found in subjects with normal length telomeres. The VUS receiving the highest pathogenic proximity score was the uncharacterized W512C variant; there was not sufficient DNA for telomere length measurement or DNA

available from other affected individuals in this family for co-segregation analysis. Of the five VUS predicted to be neutral by the pathogenic proximity score, four (H161Q, Q397E, P1107L, F1110L) did not co-segregate with disease. For comparison, no prediction method correctly classified all segregating variants, all prediction methods misclassified the two false positives, and only evolutionary conservation correctly classified the single false negative. Detailed structural hypotheses for the pathogenicity of W512C and the disease co-segregating VUS are provided in the Conclusions.

Pos	Ref	Alt	Telomere %	Segregation	PPH2	SIFT	ConSurf	PathProx	Model
55	T	S	3%	Seg	0.00	1.00	**-0.56**	-0.02	N-terminal
516	V	L	1%	Seg	0.05	0.62	**-0.15**	0.41	N-terminal
540	S	A	2%	Seg	**0.57**	0.09	**-0.80**	**0.21**	N-terminal
559	F	I	6%	Seg	**1.00**	**0.00**	**-1.11**	**0.44**	N-terminal
688	S	C	1%	Seg	**0.91**	0.14	**-0.62**	**0.27**	N-terminal
719	D	G	8%	Seg	0.03	0.22	0.21	**0.05**	N-terminal
512	W	C	Unknown	Unknown	0.17	0.48	0.31	**0.47**	N-terminal
161	H	Q	Unknown	NonSeg	0.40	0.16	**-0.35**	-0.13	N-terminal
397	Q	E	94%	NonSeg	0.08	0.20	0.40	-0.09	N-terminal
528	A	E	58%	Unknown	**0.62**	0.05	**-0.75**	**0.08**	N-terminal
574	R	W	45%	NonSeg	**0.95**	**0.00**	**-0.53**	**0.07**	N-terminal
1107	P	L	6%	NonSeg	0.63	0.01		-0.13	C-terminal
1110	F	L	Unknown	NonSeg	0	1		-0.17	C-terminal

Table 1: Variant segregation, telomeric length, and pathogenic proximity information. Variants are grouped by evidence for pathogenicity, which is inferred from disease co-segregation and patient telomere lengths. Variants that segregate with disease and short telomeres are treated as pathogenic (Figure 1). Scores in bold indicate deleterious predictions. All thresholds were applied as recommended by each method.

RTEL1 pathogenic proximity scores correlate with decreased ATPase activity in XPD mutants

RTEL1 is a RAD3-related helicase in the DEAH subfamily of the Superfamily 2 (SF2) helicases and many FIP-associated variants in RTEL1 occupy domains that are highly conserved among proteins in this family [86]. To explore the mechanistic basis for the association of RTEL1 mutations with disease, we mapped mutagenesis data from two studies of the homologous protein, XPD, onto our human model of RTEL1 (Figure 6; N=15 Fan et al.; N=10 Kuper et al.) [77,78]. Spatial proximity to pathogenic variants in RTEL1 was significantly correlated with decreased ATPase activity (Pearson r = –0.65, p = 0.0004, Figure 7a), but not with helicase activity (Pearson r = –0.36, p = 0.08, Figure 7b). This suggests that pathogenic mutations in RTEL1 may perturb ATPase activity in a manner that leads to disease. Further detailed molecular hypotheses about how the individual segregating missense variants disrupt the structure and function of RTEL1—e.g., by disrupting protein-protein interactions (W512C) or DNA binding (F559I)—are provided in the Conclusions.

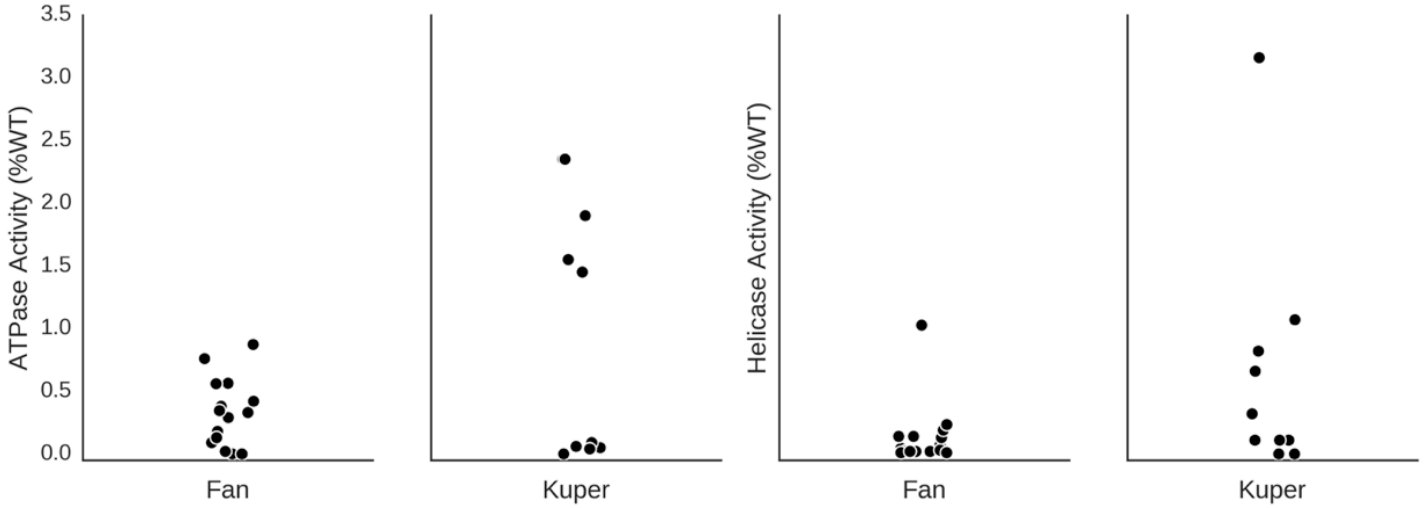

Figure 6: ATPase and helicase reported activity. Activity is reported as the percentage of wild type for missense mutations in saXPD and taXPD.

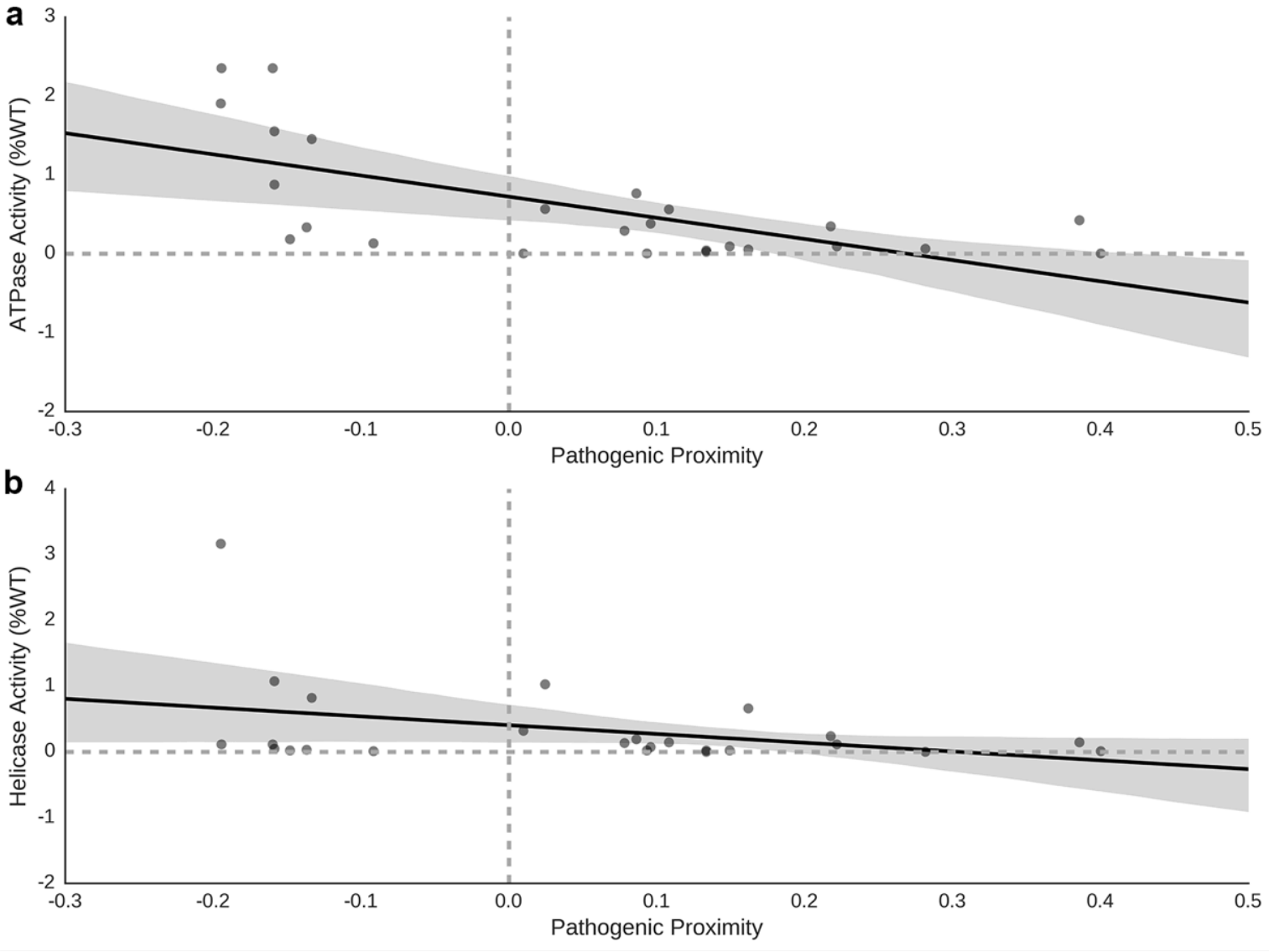

Figure 7: Pathogenic proximity scores in RTEL1 are correlated with decreased ATPase activity in mutagenesis studies of the homologous XPD protein. Pathogenic proximity scores were calculated for each missense mutation (N=25) using their position relative to known pathogenic and neutral missense variants in RTEL1. (A) Pathogenic proximity was significantly correlated with a decrease in ATPase activity (Pearson r=–0.65, p = 0.0004), but (B) not significantly correlated with changes in helicase activity (Pearson r=-0.36, p = 0.08) in the homologous XPD protein.

2.4 Conclusions

Genetic variation in *RTEL1* is a common cause of FIP in families with known genetic etiology. Most disease-causing *RTEL1* variants are private or very rare mutations and appear to reduce RTEL1 levels and/or activity[45,57]. Determining the pathogenicity of newly identified candidate VUS, particularly missense variants, presents a significant challenge in the diagnosis and treatment of patients and their family members that may be at risk[87]. Missense RVs in *RTEL1* are potentially actionable, so improved approaches to predicting pathogenicity could have a substantial clinical impact. In this report, we describe a novel, quantitative

structural approach to predicting VUS pathogenicity, applied to 13 rare missense VUS in *RTEL1*.

We constructed a homology model of the structure of RTEL1 and analyzed missense VUS relative to the spatial distribution of known pathogenic and neutral variation. Five of six VUS that segregated with FIP in families were predicted to be pathogenic by our method, as well as one VUS without disease co-segregation or telomere length data. Below, we outline potential structural mechanisms of action – ranging from disruption of protein-protein or protein-DNA interactions to destabilization of the tertiary structure of the protein – for each segregating VUS.

W512C: W512 is a bulky aromatic residue found on the surface of the structural model (Figure 8a). Surface-exposed aromatic side-chains are uncommon, and are often found to be important anchors for protein-protein binding surfaces. Replacing the tryptophan sidechain with the smaller, less hydrophobic cysteine may alter the shape and physicochemical character of a critical protein-binding surface of RTEL1, compromising its ability to perform its normal physiological function. This hypothesis is bolstered by the observation that this variant is ranked highest by our proximity score, indicating that other mutations found in close proximity to W512C – i.e. on or adjacent to the surface and likely to act through a common mechanism – are disease-linked. The importance of protein-protein interactions to RTEL1 function is underscored by the 46 unique interactions reported by the BioGrid database [88].

V516L: V516 is a moderately conserved, hydrophobic residue buried in the interior of the helicase II domain. It forms a small well-packed hydrophobic core, which lies under a patch of positively charged surface residues (R518, H713, R729, H731) (Figure 8b). Insertion of a leucine residue in this position is predicted to be destabilizing because of the additional steric bulk. Moreover, the structural rearrangement could disrupt the conformation of the basic surface patch, presumably affecting interaction with DNA.

S540A: S540 is a polar residue predicted to lie on a surface-exposed alpha helix in the helicase II domain (Figure 8c). Mutation of the hydroxyl group to an isopropyl group is predicted to have one of two effects. Either the character of the protein surface will be changed from polar to hydrophobic at that location, or, by altering the amphipathic nature of that helix, the mutation could affect the helix packing and positioning, resulting in a larger structural change such as rotation of the helix. Either of these two effects could explain the functional consequence of the variant.

F559I: F559 is a bulky aromatic residue found on the interior of the protein model, within 9 Å of the predicted DNA-binding interface (Figure 8d). Replacement of the large volume of the phenylalanine side chain with the smaller volume of isoleucine could alter the geometry of the DNA-binding cavity sufficiently to disrupt that interaction. Notably, while F559 is in the second shell of residues responsible for DNA contact, it is predicted to be directly adjacent to two first-shell residues, E591 and A621, which have been previously reported as disease-associated [59].

S688C: S688 is located on a buried helix one turn (5.9 Å) away from disease-associated residue R684. The mutation of serine to cysteine does not result in major changes in bulk, branching, charge, or hydrophobicity. However, the

presence of the sulfhydryl group in the cysteine could potentially promote misfolding and aggregation upon incorrect formation of disulfide bonds, if exposed to oxidation.

D719G: D719 is located on a surface-exposed helix near the pathogenic cluster (Figure 8e). Replacing the large charged aspartate sidechain with the single hydrogen of a glycine removes a bulky charge from the protein surface and likely disrupts the helix in that region.

T55S: T55 is a polar residue predicted to lie at the interface between alpha helices 1 and 2 (Figure 8f). Relative to the other segregating variants, T55S is distal to the pathogenic cluster and is relatively equidistant to pathogenic and neutral variation. Both threonine and serine are unusual residues to find in a helix-helix interface, and suggest that this position may be functionally important. Replacement of a threonine sidechain with that of serine does not alter the hydroxyl character of the residue, though it reduces the steric bulk by one methyl group. This is not a major volumetric change, but the removal of a beta-branching amino acid could affect inter-helical packing. This steric change could result in a relative repacking of the helix-helix interface, or could change the strength of interaction between the helices. Another mutation in this helix (K48R) has been shown to abolish ATPase activity when mutated to arginine [89], though this mutation is also physically closer to the ATP-binding cleft. Although T55 is evolutionarily conserved, SIFT and PolyPhen2 each confidently predict the serine substitution to be benign. Ultimately, there is no obvious structural basis for the pathogenicity of T55S and its distance from the pathogenic cluster suggests that any functional effects are likely impacting alternative mechanisms.

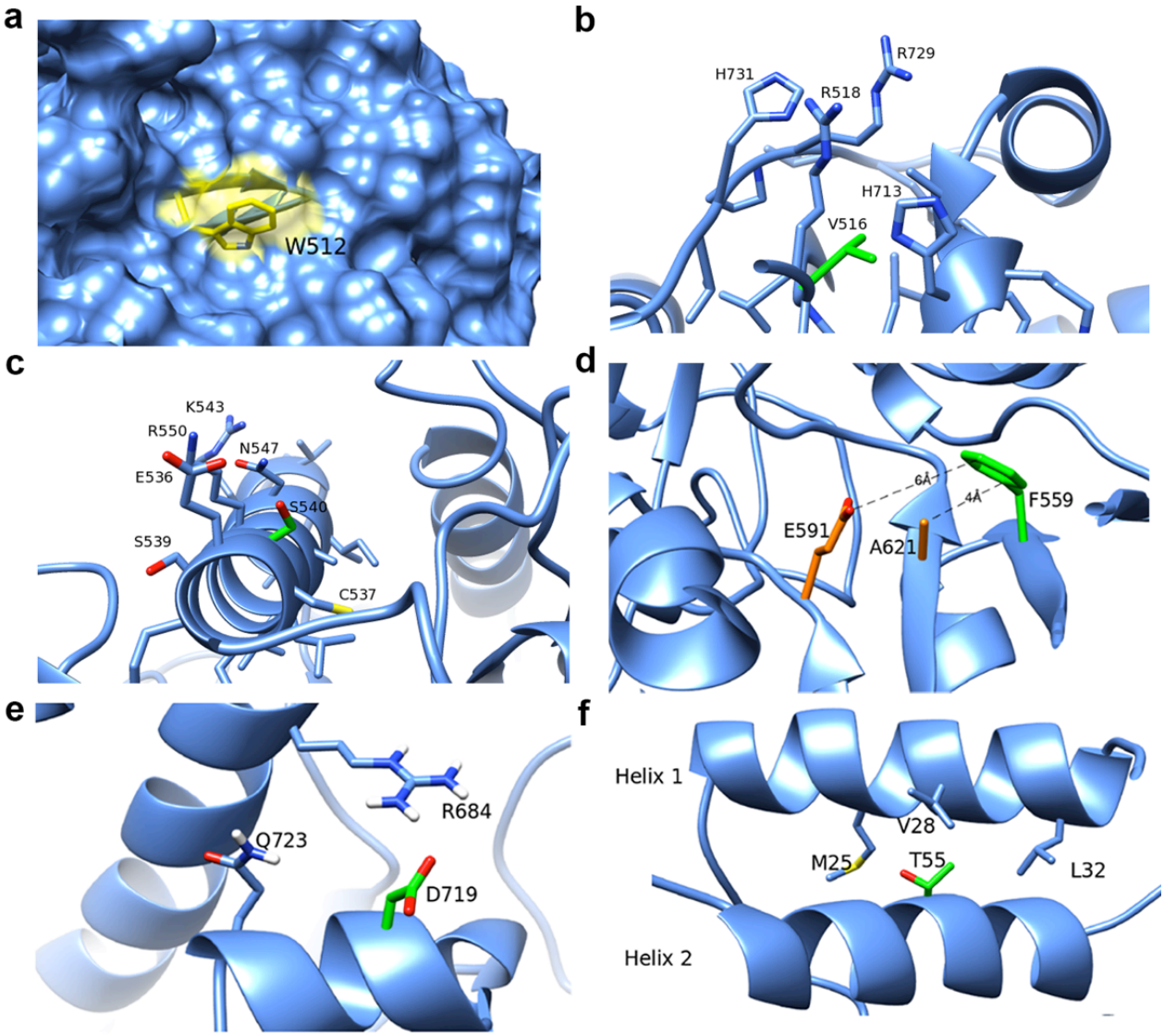

Figure 8: Structural hypotheses about the effects of six segregating *RTEL1* VUS. (a) W512 is predicted to lie on the surface of the protein. A mutation to cysteine has the potential to interfere with functionally important protein-protein interactions. (b) V516 forms a small well-packed hydrophobic core, which lies under a patch of positively charged surface residues. Mutation to leucine adds steric bulk and may induce structural rearrangements that disrupt DNA binding. (c) S540 is a polar residue predicted to lie on a surface-exposed alpha helix in the helicase II domain. Mutation to alanine may alter surface charge or cause rotation of the alpha helix. (d) F559 is buried in the core of the protein, in close proximity to residues predicted to form part of the DNA-binding cavity, including A621 and E591. Mutation to isoleucine removes steric bulk and is likely to leave a void in the hydrophobic core of the protein, disrupting structure and reducing stability. (e) D719 is predicted to fall in a surface-exposed helix. Mutation to glycine drastically reduces both the bulk and charge of the protein's surface, and likely disrupts the helix at that point. (f) T55 is predicted to form part of the interface between helices 1 and 2 in RTEL1. Mutation to a serine would reduce the steric bulk and alter the packing between the two helices.

In comparison to general pathogenicity-prediction algorithms, this approach makes use of dense population and disease-association data for variants specifically in *RTEL1* using conservative assumptions of pathogenicity. Consequently, the availability of well-characterized pathogenic and neutral variants in the protein-of-interest is essential. The incorporation of variants and

mutagenesis data from functional homologs may help to overcome this limitation. For example, the spatial distribution of disease-causing missense variants in RTEL1 suggests that the ATP-binding cleft between helicase domains I and II and the DNA-binding pore along helicase domain II are functionally critical regions of RTEL1. This finding is consistent with observed patterns of missense variants associated with *Xeroderma pigmentosum* (XP) in the homologous protein XPD [78]. While variants in XPD have different phenotypic presentations than those in RTEL1, the overlapping regions of pathogenicity suggest similar functional effects, with higher-order phenotypes driven by cellular context or unique functional domains (e.g. RTEL1 harmonin-N-like domains). This hypothesis is supported by the significant correlation between RTEL1-derived pathogenic proximity scores and reduced ATPase activity in XPD. This algorithm can be iteratively enhanced as additional disease-associated variants and primary/homologous mutagenesis data become available.

Assigning pathogenicity to missense variants in RTEL1 presents unique challenges. An ideal biomarker/assay of RTEL1 activity has not been defined, and likely differs based on the specific mutation. Short PBMC telomeres appear to be a common feature associated with RTEL1 mutations, but it is not yet clear whether this is a uniform feature; telomere length in RTEL null mouse embryonic stem cells appears stable [90], so preserved telomere length alone may not sufficiently exclude deleterious function of RTEL1 variants. In light of these complexities, for algorithm training, we conservatively defined variants as pathogenic only if they had been reported to be associated with severe pediatric disease in a recessive genetic model. For testing on novel VUS, we considered segregation with disease and telomere length in defining likely pathogenic variants. Our method classified five of the six VUS that co-segregated with FIP as pathogenic, but it also misclassified three VUS. This may demonstrate a lack of specificity when considering only the location of variants within protein structure. Spatial information demonstrates predictive potential, but it does not directly capture the impact of specific amino acid substitutions, evolutionary conservation, or biochemical information critical for interpretation. However, the specificity of our approach is comparable with other prediction methods, nearly all of which also misclassified the three VUS. It is also possible that these "misclassified" variants do adversely affect RTEL1 function without leading to a direct effect on telomere length [90]; comprehensive evaluation of these variants and others over-time should lend more clarity. At present, technical issues have limited the ability to perform in-vitro studies in overexpression systems [90]. In addition, it is possible that more than one dominant risk mutation could be found in a family; in this case, lack of co-segregation would not exclude a pathogenic effect.

Chapter 3

Quantifying spatial constraint on somatic missense mutations

The content of this chapter is adapted from a submitted manuscript: Sivley, R.M., Doux, Xiaoyi, Meiler, J., Bush, W.S., Capra, J.A., Comprehensive Analysis of Constraint on the Spatial Distribution of Missense Variants in Human Protein Structures. In revision.

3.1 Introduction

Studying the mutational landscape of tumor cells is complicated by the presence of two types of somatic mutation: those that promote the progression of cancer (driver mutations) and those that arise or persist as a result of the cancer (passenger mutations). Because passenger mutations occur within a highly dysregulated, somatic context, we expect them to experience very little selective constraint. In contrast, driver mutations must create or disrupt functional regions with tumorigenic potential; thus, driver mutations must be under selective constraint. By quantifying the spatial distribution of somatic mutations in protein structures, we can identify proteins in which somatic mutations significantly deviate from random distributions, suggesting the presence of driver mutations and identify cancer-related genes. As described in Chapter 1, we have developed a spatial analysis that tests for significant clustering (or dispersion) of missense variants within protein structure. Using this approach, we have previously demonstrated that germline disease-causing variants are spatially clustered in many protein structures. We hypothesize that somatic cancer-driving mutations exhibit similar spatial patterns due to similar contributions to disease. We further expect that clustered regions are enriched for driver mutations and will aide in the classification of driver and passenger somatic mutations.

We examined in detail four previous studies that comprehensively quantified clustering within protein structures, each of which proposed a novel methodological approach for defining and identifying spatial clusters of somatic mutations. Kamburov et al. identified clusters of nearby (<10Å) somatic mutations, weighting more heavily mutations appearing in multiple independent tumor samples, and determined significance by the extremity of the clustering score over the entire protein[22]. Tokheim et al. first identified mutations with significantly high local mutation density, and then clustered any significant hotspots within 10Å[24]. Meyer et al. first identified clusters using complete-linkage hierarchical clustering (maximum diameter of 15Å) and then measured the significance of those clusters according to their density[23]. Finally, Niu et al. filtered amino acids pairs to those at distances >20 amino acids in the primary sequence, <10Å in the protein structure, and with a significant pairwise distance within the structure, and then mapped somatic mutations into these amino acid clusters, again limiting to a somatic mutation cluster radius <10Å, measuring significance by cluster density and local recurrence rates[25]. Each of these approaches was heavily influenced by classic examples of somatic mutation clustering, and each aimed to identify slightly different patterns of somatic variation.

The discrepancy between these studies is further confounded by the use of different genetic and protein structural datasets. All studies included experimentally derived protein structures from the Protein Data Bank[91], but Tokheim et al. and Meyer et al. expanded protein coverage with the inclusion of computationally predicted homology models from ModBase[92]. Tokheim et al. and Niu et al. each used somatic mutation data from The Cancer Genome Atlas[19] (TCGA), Kamburov et al. used data from the PanCancer analysis (a subset of first dozen TCGA profiled tumor types), and Meyer et al. used somatic mutation counts from the Catalogue of Somatic Mutations in Cancer (COSMIC)[18]. These numerous methodological differences make it difficult to determine whether somatic clustering in a protein is due to real selective constraint or methodological artifacts.

Unlike previous methods, the methodology that we previously developed to analyze the spatial distribution of missense variants in protein structures is not trained or parameterized with respect to known examples of driver mutation clusters. Thus, our approach is capable of identifying clusters of somatic mutations that do not follow previously observed patterns. Furthermore, by analyzing each dataset using our consistent methodology, we facilitate direct comparison between the proteins identified by each. Finally, we evaluate the agreement between the results of our analysis and previous studies. We expect that clusters of somatic mutations identified by at least two different methods are likely true positives. Following this assumption, we compile a consensus set of genes in which significant clustering was identified by at least two spatial analyses. Our analysis is the first to analyze all previously evaluated datasets using a consistent methodology and to present the overlap in proteins identified by previous methods.

3.2 Methods

Somatic mutation and structural datasets

We analyzed somatic mutations from COSMIC version 74 and 18 cancer studies from The Cancer Genome Atlas (TCGA). COSMIC mutations were only included if they were observed in two more tumor samples to enrich the dataset for potential driver mutations. This approach was impractical for the smaller TCGA dataset, so all TCGA mutations were analyzed without regard to recurrence. Variant consequences and annotations were determined using v82 of the Ensembl Variant Effect Predictor for genomic build GRCh37[67]. Mutations were mapped into representative protein structures using Ensembl[93] transcript models, which were matched with UniProt[94] accession and Protein Data Bank[91] (PDB, 01-07-2017) IDs using cross-reference tables provided by UniProt. PDB structures were included if they were determined through x-ray crystallography or solution NMR and contained at least 20 amino acids. Reference protein sequences were aligned with observed sequences in the PDB using SIFTS[95]. Discrepancies were corrected by Needleman-Wunsch pairwise alignment with Biopython[96,97]. Computational homology models from ModBase[10] (Human 2013 and 2016) were also included to extend coverage of the proteome.

To reduce redundancy, each structural dataset was independently reduced to a minimally overlapping set of protein structures or homology models following an approach similar to Kamburov *et al.*[22]. For each structural dataset, we iteratively

selected the structure/model that provided the greatest coverage of the target protein, skipping structures with >10% sequence overlap with the existing set. For structures/models with similar sequence coverage, we selected the highest quality structure (by resolution for the PDB and the ModBase Quality Score for ModBase).

Quantifying the spatial distribution of somatic mutations in protein structure
In previous work, we developed a framework for evaluating hypotheses about the spatial distributions of genetic variants in protein structures based on Ripley's *K*, a spatial descriptive statistic commonly used in ecology and epidemiology[11,98,99]. In brief, Ripley's *K* quantifies the spatial heterogeneity of a set of variants by comparing the proportion of variants within a given distance from one another to the expected proportion under a random spatial distribution. Mutations are considered clustered if the proportion of neighbors exceeds the expectation and dispersed if the number of neighbors is lower than the expectation (Figure 9A-C). *K* is calculated across a range of distance thresholds (enabling the identification of clustering or dispersion at different scales) (Figure 9D). The distance-based scores are then summarized into a single Z-based score for each protein, determined through random permutation of mutations within the structure, where positive values indicate that mutations in the protein are clustered, and negative values indicate that mutations in the protein are dispersed (Figure 9E).

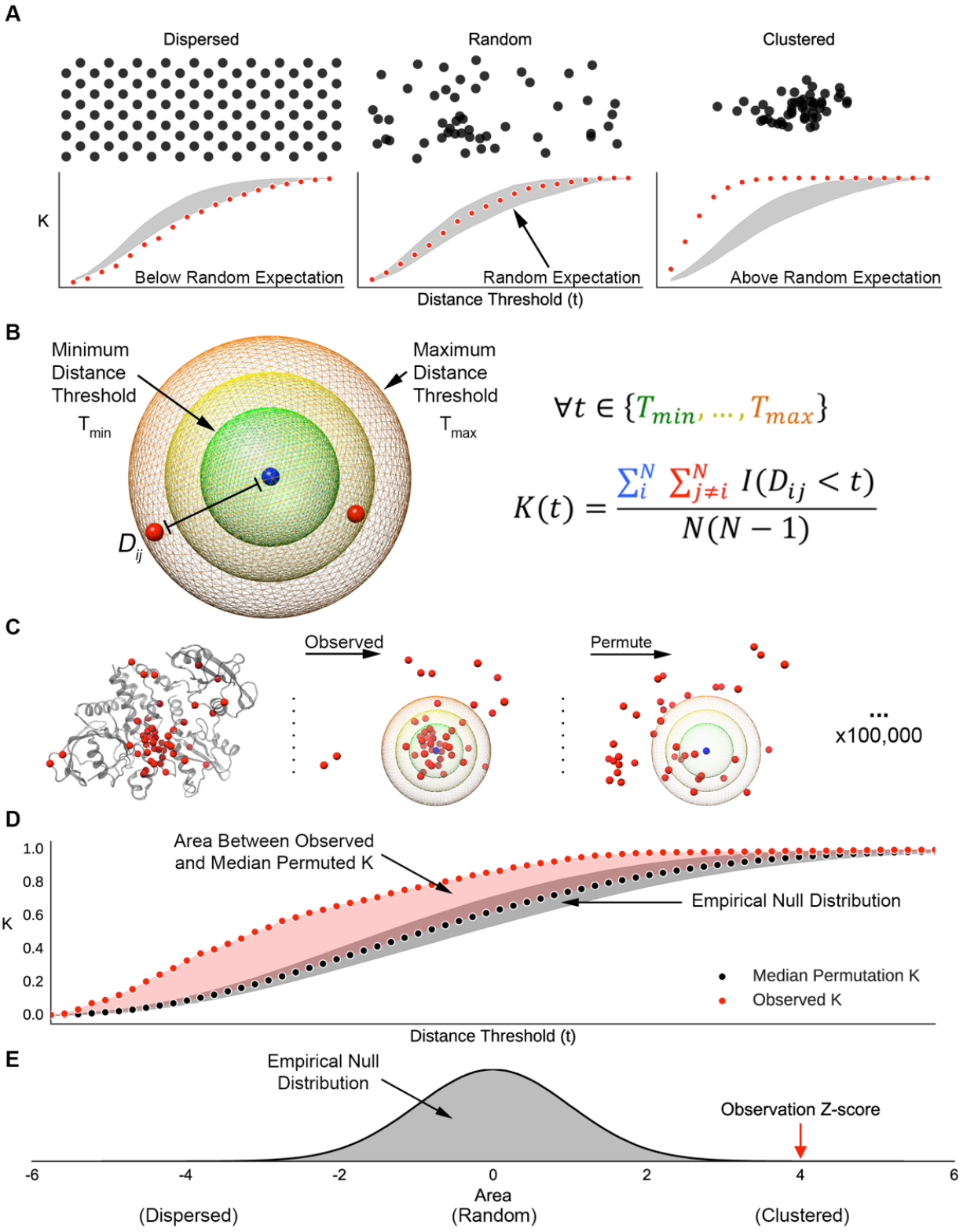
A
Dispersed
Random
Clustered
K
Below Random Expectation
Random Expectation
Above Random Expectation
Distance Threshold (t)
B
Minimum Distance Threshold
T min
Maximum Distance Threshold
T max
D ij
∀t ∈ {T min, …, T max}
K(t) = ∑i N ∑j≠i N I(Dij < t) / N(N − 1)
C
Observed
Permute
...
x100,000
D
Area Between Observed and Median Permuted K
Empirical Null Distribution
Median Permutation K
Observed K
K
1.0
0.8
0.6
0.4
0.2
0.0
Distance Threshold (t)
E
Empirical Null Distribution
Observation Z-score
-6
-4
-2
0
2
4
6
Area
(Dispersed)
(Random)
(Clustered)

Figure 9: Schematic of our framework for evaluating the spatial distribution of genetic variants. (A) Spatial distributions can diverge from random in two ways; they may have fewer neighbors than expected by chance (dispersed) or more neighbors than expected by chance (clustered). Example distributions are illustrated in reference to a random spatial distribution in 2D. Below each set of points, the resulting K statistic at multiple distance thresholds (red) is plotted in reference to the expected K distribution under a random distribution (gray). K values below the range expected at random indicate dispersion, and K values above indicate clustering. (B) Definition of the K statistic. For a range of distance thresholds (t), the number of variants neighboring each variant is computed and normalized by the total number of variant pairs. The indicator function I evaluates to 1 when two variants are neighbors (the distance between them (D_{ij}) is less than t) and 0 otherwise. (C) The observed K values are evaluated in reference to an empirical null distribution generated from 100,000 random permutations of variant locations within the protein structure. (D) The spatial distribution trend for each protein is summarized by calculating the area between the observed K values (red points) and the median permuted K values (black points). (E) This process is repeated for the K values resulting from each permuted set to generate an empirical null distribution. From this distribution, we calculate a Z-score and p-value for the observed area. Positive Z-scores indicate clustering, negative Z-scores indicate dispersion, and Z-scores near zero indicate a lack of spatial constraint.

3.3 Results

Quantifying Constraint on Spatial Patterns of Genetic Variation

We mapped genetic variants from three large variant data sets into a representative subset of 6,604 experimentally derived human protein structures from the Protein Data Bank[9] (representing 5,209 distinct proteins) and 33,144 computationally derived homology models from ModBase[100] (representing 17,984 distinct proteins). We considered the spatial distribution of 725,267 recurrent somatic missense mutations (observed in at least two human tumor samples) from the Catalogue of Somatic Mutations in Cancer[18] (COSMIC) and 31,426 somatic missense mutations from TCGA.

To evaluate the use of homology models to extend structural coverage of the proteome, we compared the COSMIC results from PDB and ModBase on shared proteins. We found that when both experimentally derived and computationally predicted structural models were available for a protein (>95% sequence overlap and excluding models for which the solved structure was used as a template, N=3,316), the spatial analysis results were highly correlated (Figure 10). Relative to the PDB, the ModBase results identified three of four significant proteins. Thus, while the analysis of computational models appears to have somewhat less power, no false positives were observed. For all analyses, we report the results on solved structures and predicted models separately. To reduce redundancy, the PDB-overlapping ModBase models were excluded from all other analyses.

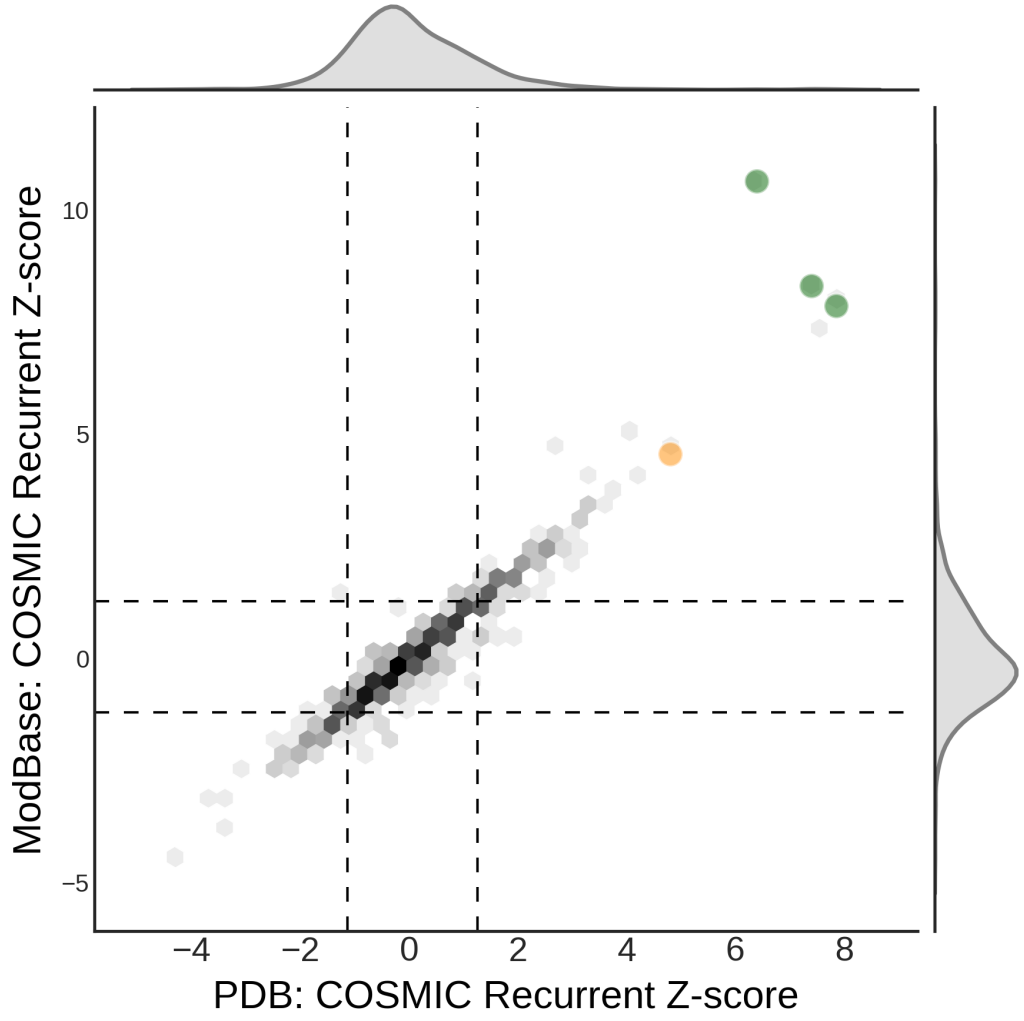

Figure 10: Spatial statistics derived from PDB structures and ModBase homology models are significantly correlated. PDB-derived spatial statistics are plotted against ModBase-derived spatial statistics on shared, sequence-matched proteins for each genetic dataset. The distribution over all pairs is shown as a density plot, with black indicating higher density. Proteins significant in the PDB analysis are shown in yellow, significant by the ModBase analysis shown in blue, and significant by both in green. We required >95% sequence overlap for each pair of PDB and ModBase structural models, and excluded any pair where the PDB structure was used as the initial template for the ModBase model.

Spatial analysis identifies clustering for both gain- and loss-of-function mutations

Early studies suggested that spatial clustering was an exclusive characteristic of oncogenes[20]; later studies have disputed the initial claim and identified somatic clusters in tumor suppressor genes as well[22–25]. Somatic driver mutations in oncogenes and tumor suppressor genes can usually be attributed to gain- and loss-of-function, respectively. Protein sequence analyses have revealed that loss-of-function variants can disrupt numerous critical elements of a protein structure, while gain-of-function variants are limited to a smaller subset of regions with functional potential[101]. We evaluated whether this relationship holds for protein structure using the dataset of dominant and recessive variants from the Human Gene Mutation Database (HGMD)[102] curated by Turner *et al.*[101]. Both dominant

and recessive variants are significantly clustered in structure (Figure 11); however, dominant variants are clustered at shorter distances (median peak significance: 8Å) than recessive variants (median peak significance: 14Å) indicating more focal clustering. The smaller clusters formed by dominant variants support the hypothesis that gain-of-function mutations are limited to specific sites with functional potential, while loss-of-function mutations more generally disrupt regions of functional importance. In summary, the frequent clustering of germline pathogenic missense variants underscores the spatial constraint on protein-coding variation and likely highlights regions of protein structures that are functionally and clinically relevant.

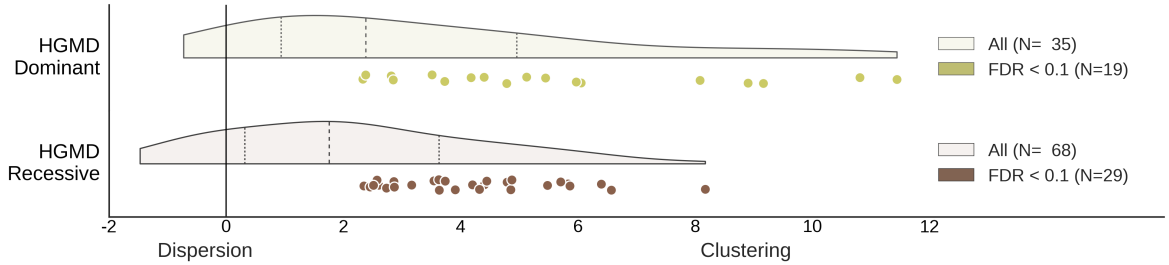

Figure 11: Autosomal dominant and recessive missense variants from the Human Gene Mutation Database (HGMD) are both spatially clustered in protein structure. However, within proteins with significantly clustered variation, dominant variants (N_{AD}=19) formed significantly smaller clusters (median peak significance distance threshold: 8Å) than recessive variants (N_{AR}=29; median peak significance: 14Å; p = 0.0005, Mann–Whitney U test). 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.

Recurrent somatic mutations are clustered in a small subset of protein structures

Several studies of tumor-derived somatic mutations have identified clustering in both sequence and structure that may highlight protein regions important for tumorigenesis[20,22–25,27]. We hypothesized that recurrent somatic mutations identified from tumor samples would exhibit patterns of spatial constraint similar to germline pathogenic missense variants. Surprisingly, we found that recurrent somatic mutations from COSMIC exhibited a weak overall trend towards spatial dispersion (Figure 12A, PDB: median Z=–0.11, ModBase: median Z=–0.12). Consistent with previous studies, we also identified significant clustering in only a small fraction of protein structures (18 of 3,080, 0.6%) and models (8 of 12,573, 0.06%).

We observed no significant difference in the overall spatial patterns between the PDB-based COSMIC and TCGA analyses (Figure 12B, PDB: median Z=-0.14, p=0.19 Wilcoxon). The TCGA analysis identified only three protein structures (of 2,884, 0.1%) with significantly clustering of somatic mutations, but these were three known cancer proteins (TP53, STK11, and PTEN), two of which were not identified by the COSMIC recurrent somatic mutation analysis (TP53 and STK11). This finding demonstrates that although we reach the same general conclusions with either somatic dataset, the specific proteins identified are not necessarily the same. This may help to explain the discrepancy between some of the previous

methods, despite all identifying proteins with significant clustering.

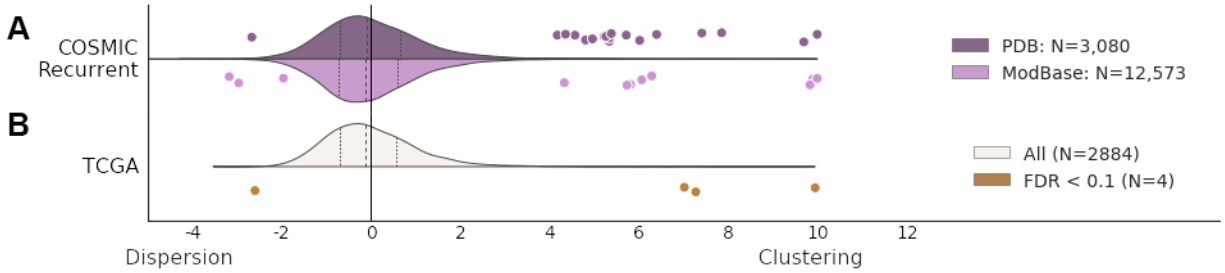

Figure 12: Distribution of spatial results for COSMIC recurrent somatic mutations and TCGA somatic mutations. Analyses are stratified by genetic dataset and the use of the Protein Data Bank or ModBase. No significant difference was observed between the overall distributions of spatial results, however the specific proteins identified as having significant clustering differ by dataset.

Overlap in significant clusters identified by different methods and datasets

The collection of 27 unique proteins we identified as containing significant clusters of somatic mutations includes many known cancer proteins[103]. 24 of the proteins we identified have been reported by at least one previous study of somatic mutation clustering[20,22–25] (Figure 13). Five proteins were identified by all methods: BRAF, EGFR, FBXW7, PIK3CA, and TP53. To our knowledge, somatic mutation clustering in the remaining 13 proteins has not been previously reported: AR, CBL, CCDC160, COMP, CREBBP, DDX3X, ITLN2, MROH2B, PCDHAC1, SEZ6, SIRPA, SMO, and TET2. Of the total 140 genes reported by any study, 32 were identified by at least two independent methods: BRAF, CDKN2A, CHEK2, EGFR, EP300, ERBB2, ERCC2, FBXW7, FGFR3, GNAS, HLA-B, HRAS, IDH1, IDH2, KEAP1, KRAS, MAP2K1, MTOR, NRAS, PIK3CA, PIK3R1, PPP2R1A, PTEN, PTPN11, RAC1, SF3B1, SMAD4, SPOP, STK11, TGFBR2, TP53, and VHL. All except HLA-B are known cancer genes included in the COSMIC cancer gene consensus. We consider these to be a high-confidence set of genes in which somatic mutations are significantly clustered in protein structure and likely to indicate clusters of driver mutations impacting functional sites important for tumorigenesis.

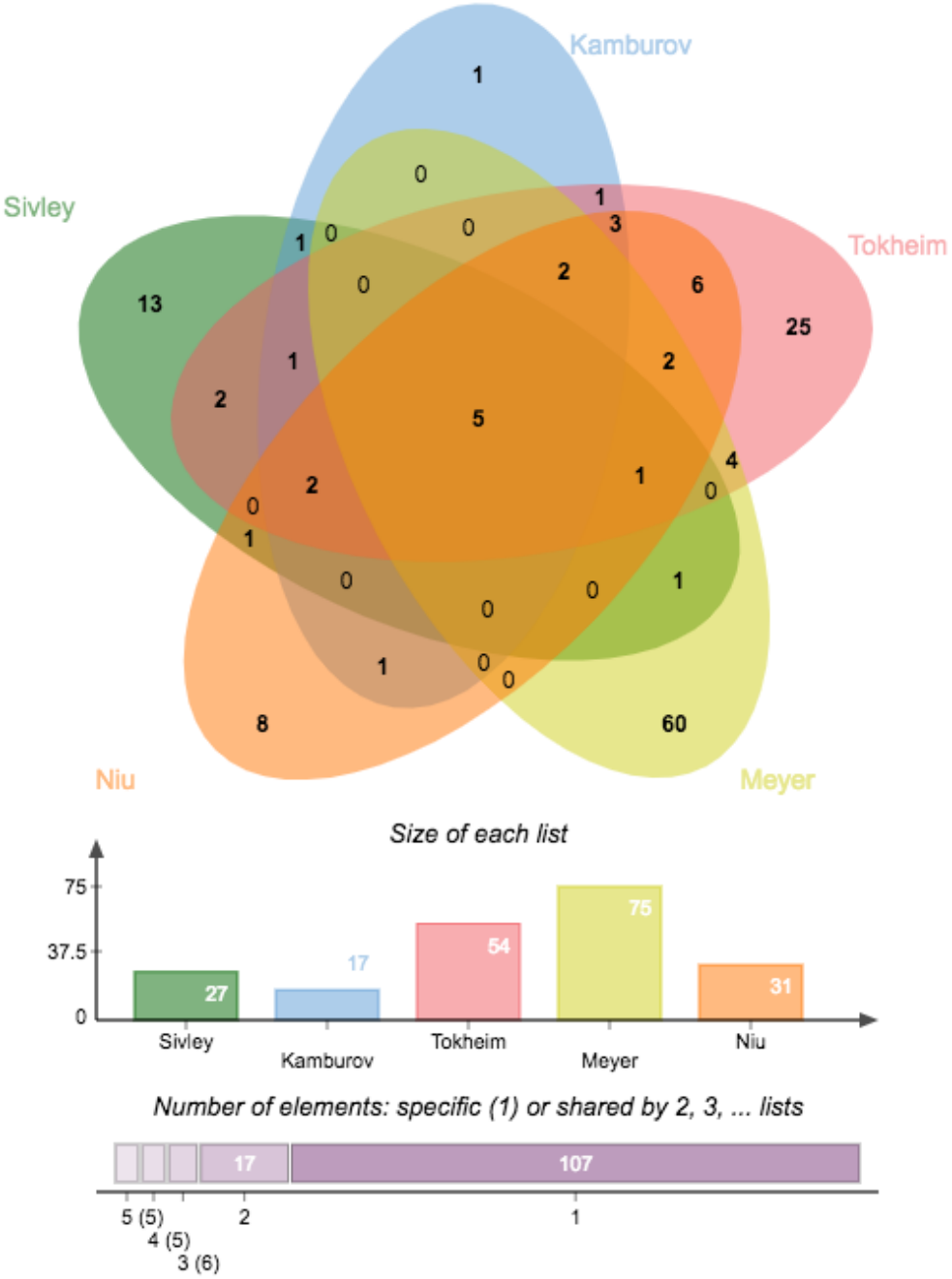

Figure 13: Proteins identified as containing significant clustering of somatic mutations. We compared the results of our comprehensive, uniform analysis of COSMIC and TCGA distributions in PDB structures and ModBase models to previous studies of somatic mutations in cancer. There is a large variance in the number of proteins identified by different studies, but there is also substantial overlap.

Spatial clustering re-identifies functional clusters of driver mutations: PTPN11

An important assumption of this analysis is that significant clustering of somatic mutations is indicative of driver mutations densely affecting a functional site important for tumorigenesis. We present here one example of a cancer-relevant functional site identified by our data-driven spatial analysis. Recurrent somatic mutations in PTPN11 [MIM:

176876], which encodes the protein tyrosine-protein phosphatase non-receptor type 11 (SHP-2), are clustered at the structural interface between the protein tyrosine phosphatase (PTP) and Src-homology 2 (SH2) domains (Figure 14). Germline pathogenic missense variants at this interface are associated with LEOPARD syndrome (LPRD1 [MIM: 151100]), Noonan syndrome (NS1 [MIM: 163950]), and increased risk for juvenile myelomonocytic leukemia (JMML [MIM: 607785]). Somatic mutations to PTPN11 are often found in leukemias and several solid tumors[104]. The relative orientation of the PTP and SH2 domains determines whether SHP-2 is in its active or inactive state. Disease-causing mutations have been shown to disrupt the interaction interface between these domains, with mutations causing NS1 leading to a more energetically favorable active state relative to wild-type[105] (gain-of-function) and mutations causing LPRD1 resulting in an inactive state[106] (dominant negative). Our analysis identified significant spatial clustering of somatic mutations in several known cancer genes. This example demonstrates how these clusters can identify functional sites within protein structures and help to elucidate the structural mechanisms driving tumorigenic effects.

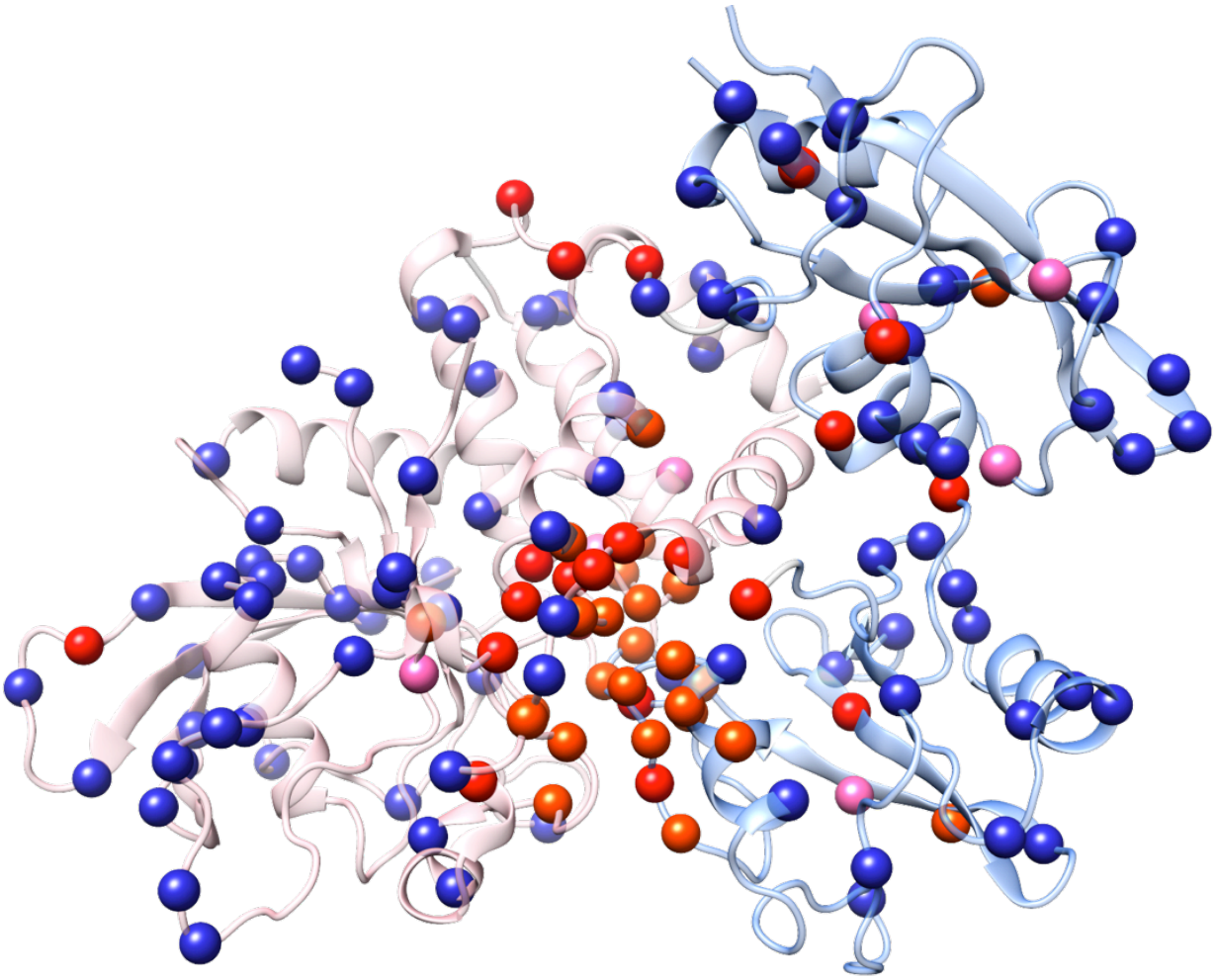

Figure 14: Distribution of COSMIC recurrent somatic mutations in SHP-2. Our analysis identified significant clustering of *PTPN11* somatic mutations in the structure of its protein, SHP-2. Mutations were primarily located at the structural interface of the PTP and SH2 domains, and impact the affinity of the two domains. Increase and decreased binding affinity at this interface each lead to distinct clinical phenotypes. This example demonstrates how the identification of somatic mutation clusters can help to identify tumorigenic sites and the molecular basis of driver mutations.

Analysis of protein structure reveals significant patterns of spatial constraint not identified from protein sequence

Experimentally derived protein structures are available for approximately 22% of human proteins. Computationally derived homology models expand coverage (of at least part of the protein) to 77%, but there are thousands of human proteins for which we do not have reliable structural information. The linear protein sequence is available for all proteins, but does not represent the functional context of the protein. Thus, we hypothesized that significant spatial patterns within the three-dimensional protein structure may not be identifiable from protein sequence alone. We repeated our analysis using the protein sequence of each experimentally derived protein structure to compute the linear K statistic and measured the overall

correlation and predictive performance compared to structure-based K analyses. There is little overlap in the proteins identified as significantly constrained by each analysis. Sequence-based analyses of missense variation identified only 37% of the significant spatial patterns identified in protein structure, suggesting that many significant spatial patterns in protein structure are introduced by protein folding. Conversely, the sequence analysis, relative to the structural analysis, had a precision of 0.58, which indicates that significant spatial patterns of variants in protein sequence are often disrupted in the folded protein structure. Overall, the statistics for sequence and structure are correlated (Spearman's rho=0.37 $p=9.6x10^{-101}$; Figure 15), but proteins without significant constraint in either sequence or structure drive this pattern. These results demonstrate that sequence-based analyses do not accurately predict significant spatial constraint on missense variation in protein structure.

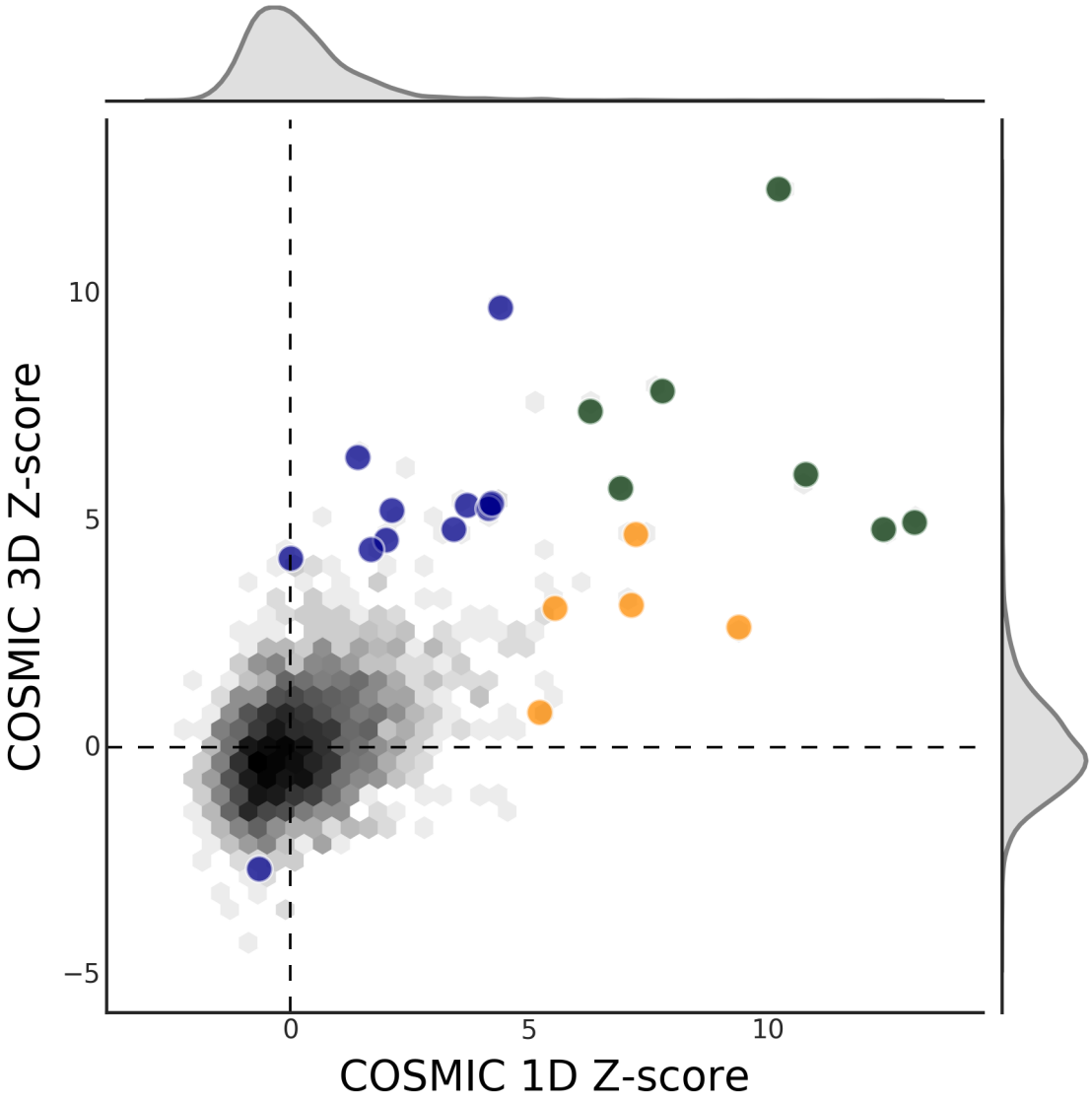

Figure 15: Protein sequence is a poor predictor of spatial patterns in protein structure. The Ripley's K Z-score for significant spatial constraint on each protein in the PDB set computed over its 3D structure is contrasted with the K Z-score computed using its 1D sequence. The distribution over all structures is shown as a density plot, with black indicating higher density. Large circles indicate structures with spatial distributions significantly different from random; circles are colored blue if significant in the structural analysis, yellow if significant in the sequence analysis, and green if significant in both analyses. The sequence- and structure-derived Z-scores are (Spearman's rho=0.37 $p=9.6x10^{-101}$), but sequence analysis identified very few proteins with significant spatial distributions in protein structure.

3.4 Conclusion

By projecting hundreds of thousands of somatic mutations observed in human tumors into three-dimensional protein structures, we comprehensively quantified patterns of spatial constraint on human somatic mutations within their functional and structural context. In contrast to the strong and consistent clustering of germline pathogenic missense variation observed in our previous work, significant clustering of recurrent somatic mutations was identified in relatively few proteins.

The stronger clustering of germline disease-causing variation compared to recurrent somatic variants may reflect differences in spatial constraint and phenotypic effects of variation outside of the germline[107]. There are likely differences in variant tolerance between germline and somatic contexts; germline variants are present in all tissues and are subject to many powerful constraints throughout development. In contrast, somatic variants influence only a subset of tissues and developmental time points, and thus may be tolerated in contexts that would be lethal in the germline. Alternatively, germline and somatic differences may be attributable to relaxed constraint within the tumor context, which is already highly dysregulated. Somatic datasets also likely contain many unconstrained, neutral passenger mutations, which may further explain the spatial randomness of somatic mutations across most protein structures.

Several studies have examined the spatial clustering of somatic mutations within protein structures. The number of proteins exhibiting somatic mutation clustering varies greatly between studies: Kamburov *et al.* identified only 17 proteins with significant somatic clustering while Meyer *et al.* identified 75 proteins with high-scoring somatic clusters. Our analysis identified 27 proteins with significantly clustered recurrent somatic mutations, of which 24 had been previously identified. The variation between methods is attributable to differences in many aspects of the studies, including clustering algorithms, mutation cluster definitions, limits on cluster size, and the genetic and structural datasets considered. Prior approaches have also focused on the identification of *clusters* of somatic variants and were parameterized using known examples in the cancer literature, which may not be applicable in a germline context and cannot capture spatial dispersion. Key advances of our approach to characterizing spatial distributions are that it identifies both significant clustering and dispersion (at any distance) compared to a random distribution and makes no domain-specific assumptions. As a result, our method captures additional patterns of spatial constraint of genetic variation over all proteins. This may consequently reduce its power to identify some somatic mutation clusters detected by cancer-targeted approaches, in particular those that detect clusters of one or two highly recurrent mutations. However, we note that our method identifies a similar number of proteins as Kamburov *et al*., who similarly aimed to identify proteins with significant overall clustering of somatic mutations. This may suggest that the large number of proteins identified by other methods is largely attributable to disagreement as to what constitutes somatic mutation clustering, rather than better power to detect clusters that we failed to identify.

The selection of mutation datasets also influences the power of different methods to detect spatial patterns. COSMIC is a submission-based database of somatic mutations, and maximizes the number of available variants for analysis. However, the use of a submission-based system introduces the potential for reporting bias into the representation of proteins and mutations. In contrast, the Cancer Genome Atlas (TCGA) provides consistent, whole-exome sequencing data from many cancer studies and tumor types, but has much smaller sample size; data from 18 TCGA studies did not include enough *recurrent* mutations to satisfy our inclusion criteria. COSMIC identified significant clustering in far more proteins than TCGA, and we observed no significant difference in the overall distribution

of COSMIC and TCGA results, suggesting that bias in the COSMIC dataset improved our power, but did not critically affect our general findings.

Chapter 4

Identifying the Clinical Impact of Loss-of-Function Intolerant Genes using PheWAS

4.1 Introduction

We have thus far focused our analyses of selective constraint on refining disease-associations within the context of protein structure. However, we can likewise use selective constraint to guide the discovery of novel phenotypic associations. Large-scale efforts to whole-exome sequence tens of thousands of individuals has provided a wealth of information about where genetic variation is and is not tolerated throughout the human genome[7]. This information led to the identification of over three thousand genes with evidence of loss-of-function (LoF) intolerance (LoFi). The high degree of constraint on variation within these genes suggests critical importance to human health, yet most are not associated with any human disease. Using constraint as a guide, we explore the phenotypic impact of protein-coding variation in LoFi genes to elucidate their function and clinical relevance.

Using phenotypic information derived from the Vanderbilt EMR and dense exome genotyping from Vanderbilt's BioVU biobank, we performed a targeted gene-level phenome-wide association study (PheWAS) to characterize the clinical impact of genetic variation within LoF-intolerant (LoFi) genes. By definition, we do not expect to find many (if any) LoF variants within most of these genes. However, we expect that within a gene, single-nucleotide variants (SNVs) and LoF variants are associated with similar phenotypes [31]. Because we're interested in characterizing entire genes, we use gene-level sequence-kernel association test (SKAT-O) that aggregate information across all SNVs within each gene for association with each phenotype. Using this approach, we interrogate LoFi genes for association with clinical phenotypes and contrast the clinical and mammalian phenotypes associated with loss-of-function tolerant and intolerant genes.

4.2 Methods

Dataset curation and quality control

The dataset used in the analyses described were obtained from Vanderbilt University Medical Center's BioVU, which is supported by institutional funding and by the CTSA grant ULTR000445 from NCATS/NIH. Genome-wide genotyping was funded by NIH grants RC2GM092618 from NIGMS/OD and U01HG004603 from NHGRI/NIGMS. Consent for participation in the BioVU resource during the ascertainment period for data included in this study was opt-out. We identified 26,577 samples with dense exome genotyping on the Illumina HumanExome BeadChip v1.0. These samples were ascertained from six partially overlapping patient cohorts: pediatrics, elderly, cancer, rare phenotypes, longitudinal, and samples included in previous GWAS. Presence or absence within each of these cohorts was included as covariates in all association tests. We removed all samples with genotyping efficiency less than 95%, followed by the

removal of all remaining variants with genotyping efficiency less than 95%. We next identified and removed duplicate variants. We then identified and removed any samples with estimated proportion of identify by descent greater than 0.2 and all samples with gender inconsistencies.

To reduce confounding by continental ancestry, we filtered the dataset to samples of predominantly European ancestry. We first merged the BioVU dataset with samples from the 1000 Genomes[21] and performed principal components analysis (PCA) using ancestry-informative markers available on the Illumina HumanExome BeadChip. Using model based clustering—trained on 1000 Genomes continental ancestry—we assigned continental ancestry to each BioVU sample[108] and retained only those samples classified as having European ancestry. We then performed a second iteration of PCA using only the BioVU samples of European ancestry with all 1000 Genomes samples, and selected the top five principal components for use as covariates in all association tests.

To identify variants in the protein-coding regions of genes, we processed the BioVU European-ancestry dataset with the Ensembl Variant Effect Predictor[67] (VEP) and reduced the dataset to missense, synonymous, and protein-truncating variants. The gene annotation provided by VEP was used to group variants for all association tests. LoF-intolerant genes with fewer than two protein-coding variants were excluded from analysis. In summary, we tested 173,385 protein-coding variants in 2,457 genes for association with 1,480 EMR-derived phenotypes in 21,388 clinical samples of predominantly European ancestry; not all samples were included in all association tests (phenotype-specific exclusions, gender-specificity, etc).

Gene-level rare variant PheWAS for EMR-derived phenotypes

Gene-level, rare-variant (maximum MAF < 0.05) association tests were performed using the optimized sequence-kernel association test[32,33] (SKAT-O). We required a minimum of 20 cases for a phenotype to be included in the analysis; this criteria was met by 1,480 of 1,816 phenotypes. BioVU is an EMR-derived clinical cohort, and thus most association tests for any particular phenotype will have a large imbalance between cases and controls; additionally, all variants included in the analysis are low frequency. To account for these aspects of the dataset, we employ the (default) SKAT-O hybrid method for calculating p-values , which selects a p-value correction method (no adjustment, efficient resampling (ER), adaptive ER, moment matching adjustment, or quantile adjusted moment matching) on the basis of minor allele count, the number of individuals with the minor allele, and the degree of case-control imbalance [109]. For each association test, we first construct a null model containing age, sex, ascertainment cohort, and the first five principal components. We then run SKAT-O on the binary PheWAS code for each EMR-derived phenotype. Finally, we applied a false discovery rate (FDR) threshold of 1% to account for multiple testing.

Validating statistically significant results with mouse knockout phenotypes

To determine if biological support was available for significant associations with

ICD9-derived phenotypes, we merged our results with data from the Mouse Genome Informatics[110] (MGI) Human-Mouse Disease Connection (HMDC), which links mouse model phenotypes to homologous human genes Significant associations were considered biologically supported if the PheWAS category associated with the ICD9-derived phenotype matched semantically with any of the top-level mammalian phenotype categories assigned to the associated gene. This is not intended to replace biological validation, nor is it intended to classify unsupported associations as spurious. Rather, data from the HMDC is intended to assess the proportion of significant clinical associations with existing evidence for biological support and the highest likelihood of biological validation.

Clinical phenotype enrichment amongst LoFi genes
Each ICD9-derived phenotype is assigned to a PheWAS category; for example, rheumatoid arthritis is assigned to the musculoskeletal category. To determine if certain PheWAS categories were enriched for significant associations with LoFi genes, we constructed a 2x2 contingency table of all association tests by significant ($q<0.1$)/not significant ($q>=0.1$) and annotated/annotated with each PheWAS category. Odds ratios represent the enrichment or depletion of statistically significant associations within each PheWAS category, exclusively among LoFi genes. To determine if these enrichments were specific to LoFi genes, we repeated the analysis using all other genes. To determine if significantly-associated phenotype categories were significantly enriched for LoFi genes relative to all other genes, we constructed a 2x2 contingency table of significant associations by loss-of-function tolerance/intolerance and annotated/not annotated with each PheWAS category; odds ratios represent the enrichment or depletion for significantly associated LoFi genes within each PheWAS category.

4.3 Results

Dataset quality control and continental ancestry determination
Using samples from the 1000 Genomes Project[21], we performed principal components and model-based clustering to determine the predominant continental ancestry of each BioVU sample (Figure 16A-C). We then filtered the dataset to the 21,388 samples classified with predominantly European ancestry, and repeated the analysis to generate principal components to be used as covariates during association testing (Figure 16D). After quality control and ancestry filtering, the dataset included 21,388 samples with genotyping data for 173,385 protein-coding variants. In total 2,457 LoFi genes and 11,916 non-LoFi genes included at least two genotyped rare variants.

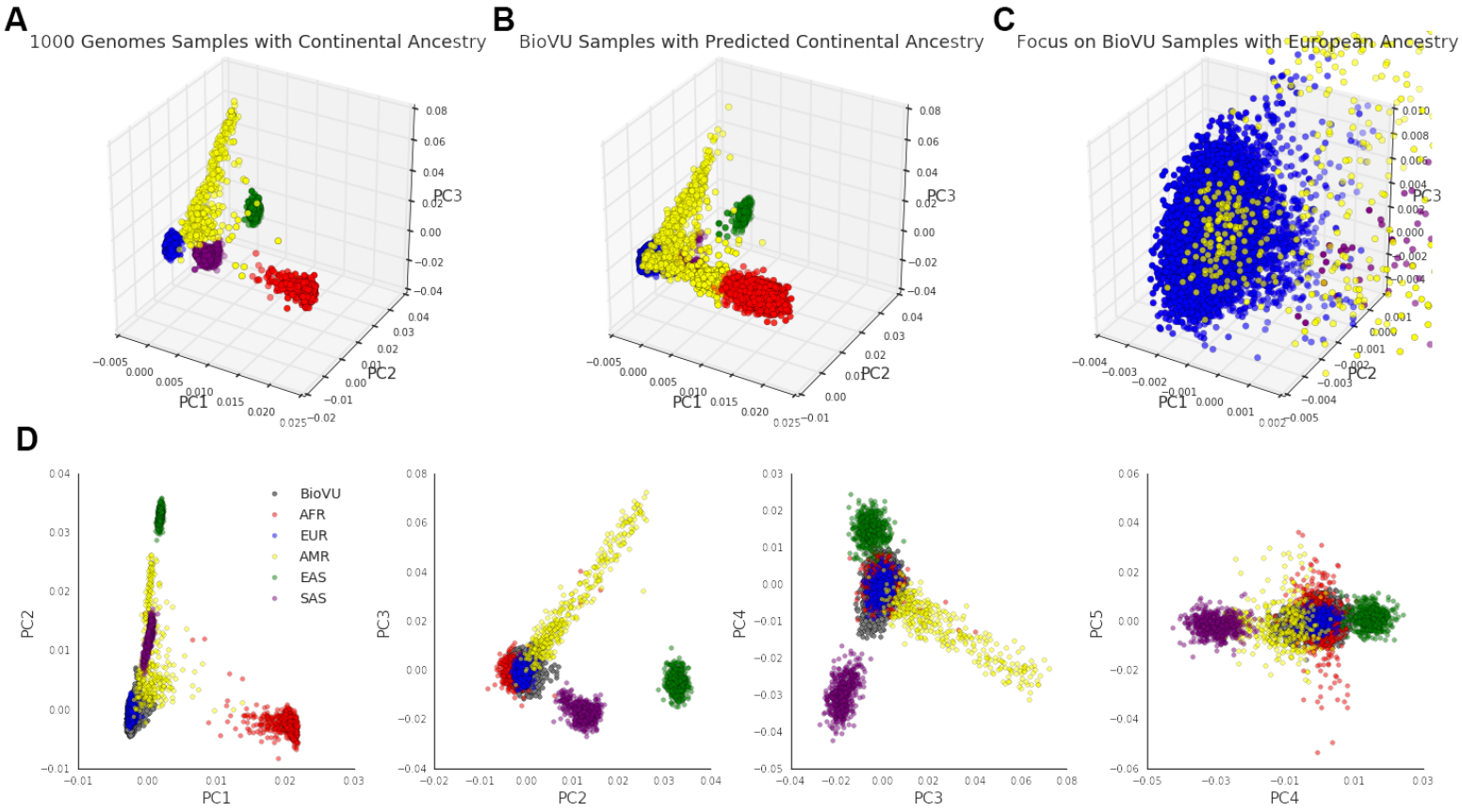

Figure 16: BioVU genetic ancestry assignment and principal components analysis. Principal components analysis (PCA) was performed using all BioVU samples passing quality control along with samples from the 1000 Genomes Project. (A) The first three principal components are plotted for ancestry-labeled 1000 Genomes samples and (B,C) ancestry-assigned BioVU samples. (D) A second iteration of PCA was performed using only the BioVU samples with predominantly European ancestry along with samples from the 1000 Genomes; the top five principal components were included as covariates in all association analyses.

Replication rate of known gene-phenotype associations

To assess the ability of our approach to identify gene-level associations with clinical phenotypes, we first determined the proportion of all genes with known gene-phenotype associations for which we identified significant associations in our comprehensive analysis (not limited to LoFi genes). In total, we identified significant associations for 164 (of 3,073, ~5%) genes with known gene-phenotype associations from OMIM. These included many gold standard examples of allelic heterogeneity, including the association of CFTR with cystic fibrosis (CF, N_{cases}=116, p=1.70x10^{-10}, SKAT-O) and of PAH with phenylketonuria (PKU, N_{cases}=23, p=4.43x10^{-26}, SKAT-O). However, other well-characterized phenotypes were notably absent; for example, we did not detect a significant association between HGD and alkaptonuria (AKU, N_{cases}=43, p=0.52, SKAT-O). This may reflect a lack of specificity in the ICD9 code for AKU (270.2), which captures all disturbances of aromatic amino-acid metabolism, including clinically distinct phenotypes like albinism and Waardenburg syndrome. Alternatively, it may indicate an overuse of more general ICD9 codes, as evidenced by the 260 samples

billed with ICD9 270, which encompasses all disorders of amino acid transport and metabolism. These findings suggest that ICD9-derived are capable of replicating known phenotypic associations, but may lack sensitivity for phenotypes that are not well captured by billing codes.

Amongst LoFi genes with known gene-phenotype associations in OMIM, 39 (of 625, ~6%) were significantly associated with clinical phenotypes in our analysis. This was only slightly higher than the proportion of genes without previous association to disease significantly associated with clinical phenotypes (81 of 1,832, ~4%). In summary, the 81 LoFi genes with significant associations to clinical phenotypes likely represent a small proportion of the phenotypic impact of LoFi genes, but nonetheless provide novel insights about their phenotypic effects.

Rare variant PheWAS of LoFi genes identifies significant phenotypic associations

We performed a rare-variant PheWAS using SKAT-O and 1,480 ICD9-derived clinical phenotypes for 2,457 genes predicted to be loss-of-function intolerant (LoFi)[7]. We identified 129 gene-phenotype associations significant at a false discovery rate (FDR) of 1% (Table 2, Figure 17A) and replicated several known, unambiguous gene-phenotype associations. For example, rare variants in *SERPINA1* [MIM 107400], which encodes the protein Alpha-1 antitrypsin, were significantly associated with Alpha-1 antitrypsin deficiency ($p=5.30x10-^{272}$, A1ATD [MIM 613490]). The most significant associations were four genes associated with rheumatoid arthritis (RA [MIM 180300]): PTPRB ($p=3.23x10^{-37}$, DMXL1 ($p=1.45x10^{-35}$), KRT6A ($p=8.54x10^{-34}$), and PCLO ($p=2.32x10^{-30}$), none of which have been previously associated with RA. To test for enrichment for specific phenotypes amongst LoFi genes, we performed association tests for all non-LoFi genes meeting the inclusion criteria as well (Figure 17B).

Gene Name	PheWAS Category	PheWAS Description	Cases	P-value
PTPRB	musculoskeletal	Rheumatoid arthritis	836	3.23E-37
DMXL1	musculoskeletal	Rheumatoid arthritis	836	1.48E-35
KRT6A	musculoskeletal	Rheumatoid arthritis	836	8.54E-34
PCLO	musculoskeletal	Rheumatoid arthritis	836	2.32E-30
PTPRB	musculoskeletal	Rheumatoid arthritis and other inflammatory polyarthropathies	1066	1.24E-28
DMXL1	musculoskeletal	Rheumatoid arthritis and other inflammatory polyarthropathies	1066	4.03E-27
KRT6A	musculoskeletal	Rheumatoid arthritis and other inflammatory polyarthropathies	1066	5.02E-27
PCLO	musculoskeletal	Rheumatoid arthritis and other inflammatory polyarthropathies	1066	1.67E-22
NUMA1	musculoskeletal	Cyst of bone	22	9.74E-11
CLUH	musculoskeletal	Other and unspecified disc disorder	36	6.47E-09
KIAA1468	musculoskeletal	Contracture of tendon (sheath)	69	1.22E-08
PIKFYVE	musculoskeletal	Contracture of tendon (sheath)	69	2.88E-08
KMT2A	musculoskeletal	Arthropathy associated with other disorders classified elsewhere	85	6.23E-08
CHAMP1	musculoskeletal	Pathologic fracture of femur	58	1.34E-07
GIGYF2	musculoskeletal	Other and unspecified disorders of back	154	2.01E-07
SNAP91	musculoskeletal	Flat foot	58	2.45E-07
CRAMP1L	musculoskeletal	Kyphosis (acquired)	77	3.08E-07

BCR	musculoskeletal	Juvenile osteochondrosis	20	3.38E-07
SLC4A4	endocrine/metabolic	Alpha-1-antitrypsin deficiency	21	6.35E-22
PRR14L	endocrine/metabolic	Alpha-1-antitrypsin deficiency	21	5.42E-16
DST	endocrine/metabolic	Lipoprotein disorders	24	2.01E-11
CD72	endocrine/metabolic	Pituitary hyperfunction	47	3.84E-10
SSH2	endocrine/metabolic	Cushing's syndrome	40	8.45E-10
SUPT16H	endocrine/metabolic	Disorders of urea cycle metabolism	25	6.47E-09
ARHGAP31	endocrine/metabolic	Morbid obesity	822	1.03E-08
IGF2R	endocrine/metabolic	Other immunological findings	221	1.23E-08
USP47	endocrine/metabolic	Autoimmune disease NEC	26	3.09E-08
PHC3	endocrine/metabolic	Nonspecific abnormal results of other endocrine function study	50	4.93E-08
CDC5L	endocrine/metabolic	Carcinoid syndrome	32	6.45E-08
DST	endocrine/metabolic	Other disorders of lipoid metabolism	37	6.33E-08
CAMK1D	endocrine/metabolic	Deficiency of humoral immunity	57	1.38E-07
NRP1	endocrine/metabolic	Nonspecific abnormal results of other endocrine function study	50	2.53E-07
KANSL3	endocrine/metabolic	Abnormal results of function study of thyroid	70	3.41E-07
NUP98	mental disorders	Acute reaction to stress	41	1.25E-13
BMP7	mental disorders	Acute reaction to stress	41	1.81E-09
CELSR3	mental disorders	Mental disorders durring/after pregnancy	31	2.12E-08
PCNX	mental disorders	Decreased libido	26	2.40E-08
LMNB1	mental disorders	Symptoms involving head and neck	25	2.91E-08
RAP1GAP	mental disorders	Symptoms involving head and neck	25	4.78E-08
MAML2	mental disorders	Paranoid disorders	28	5.51E-08
SHANK2	mental disorders	Vascular dementia	81	5.67E-08
MON2	mental disorders	Mental disorders durring/after pregnancy	31	9.09E-08
FST	mental disorders	Aphasia/speech disturbance	433	1.14E-07
MAP7	mental disorders	Paranoid disorders	28	2.66E-07
HTT	infectious diseases	Sexually transmitted infections (not HIV or hepatitis)	24	5.05E-13
FMN2	infectious diseases	H. pylori	32	7.76E-11
RP11-1055B8.7	infectious diseases	Infestation (lice, mites)	31	3.08E-10
DNMT3B	infectious diseases	H. pylori	32	6.44E-09
PAX3	infectious diseases	Sexually transmitted infections (not HIV or hepatitis)	24	6.42E-08
UST	infectious diseases	Viral hepatitis B	75	1.06E-07
PRDM1	infectious diseases	Sexually transmitted infections (not HIV or hepatitis)	24	1.53E-07
ACIN1	circulatory system	Aneurysm and dissection of heart	28	5.81E-12
ERBB2	circulatory system	Polyarteritis nodosa	27	6.01E-10
PHLPP1	circulatory system	Mobitz II AV block	20	6.81E-10
PRR12	pregnancy complications	Infections of genitourinary tract during pregnancy	24	7.77E-12
GPHN	pregnancy complications	Excessive vomiting in pregnancy	26	7.72E-11
PML	pregnancy complications	Missed abortion/Hydatidiform mole	39	1.14E-08
RELN	pregnancy complications	Miscarriage; stillbirth	91	4.40E-08
LPHN3	pregnancy complications	Early onset of delivery	51	1.01E-07

NEURL4	pregnancy complications	Other complications of pregnancy NEC	51	1.58E-07
PLEKHO1	pregnancy complications	Interstitial emphysema and related conditions of newborn	20	2.93E-07
SCUBE1	neoplasms	Benign neoplasm of other female genital organs	24	1.60E-11
SEMA6A	neoplasms	Benign neoplasm of other female genital organs	24	4.18E-08
CBX2	neoplasms	Cancer of the gums	37	8.60E-08
PRKCQ	neoplasms	Radiotherapy	453	1.38E-07
PTPRT	neoplasms	Bone marrow or stem cell transplant	21	2.58E-07
PRDM2	neoplasms	Benign neoplasm of eye, uveal	55	3.14E-07
AKAP8	neoplasms	Cancer of major salivary glands	71	3.25E-07
PIKFYVE	congenital anomalies	Other congenital anomalies of lower limb, including pelvic girdle	56	5.44E-11
PIKFYVE	congenital anomalies	Congenital hip dysplasia and deformity	50	1.99E-10
NUP85	congenital anomalies	Obstructive genitourinary defect	65	1.42E-08
AHDC1	congenital anomalies	Congenital anomalies of female genital organs	23	3.18E-07
USP19	digestive	Anomalies of tooth position/malocclusion	88	6.13E-11
USP19	digestive	Dentofacial anomalies, including malocclusion	97	4.94E-10
TRAPPC8	digestive	Hepatomegaly	68	5.08E-10
HTR1A	digestive	Anomalies of jaw size/symmetry	26	2.86E-08
ZC3H13	digestive	Leukoplakia of oral mucosa	42	6.74E-08
ARHGAP31	digestive	Bariatric surgery	163	2.50E-07
ZNF609	injuries & poisonings	Complication of amputation stump	40	7.54E-11
KIF1B	injuries & poisonings	Muscle/tendon sprain	37	5.03E-10
BCAS3	injuries & poisonings	Subarachnoid hemorrhage (injury)	61	2.83E-09
PRR12	injuries & poisonings	Poisoning by water, mineral, and uric acid metabolism drugs	57	2.47E-08
PLXNC1	injuries & poisonings	Spinal cord injury without evidence of spinal bone injury	22	2.66E-08
RAVER1	injuries & poisonings	Subarachnoid hemorrhage (injury)	61	5.99E-08
MIER2	injuries & poisonings	Postoperative shock	39	1.12E-07
DGKZ	injuries & poisonings	Adverse effects of sedatives or other central nervous system depressants and anesthetics	59	1.29E-07
ZNF407	injuries & poisonings	Open wound of foot except toe(s) alone	41	3.16E-07
NOL6	symptoms	Nonallopathic lesions NEC	27	8.72E-11
SCYL2	symptoms	Abnormal posture	118	2.18E-10
PCNX	symptoms	Symptoms of the muscles	120	1.64E-08
HOXC6	symptoms	Cramp of limb	62	2.93E-08
PLK1	symptoms	Rhabdomyolysis	45	3.41E-08
CBX2	symptoms	Elevated carcinoembryonic antigen [CEA]	93	3.15E-07
TNFAIP3	neurological	Parasomnia	21	2.16E-10
YTHDC2	neurological	Cerebral cysts	36	2.37E-09
C18orf25	neurological	Trigeminal nerve disorders [CN5]	130	3.80E-08
FAM208B	neurological	Nerve root lesions	58	3.54E-07
FCHO1	respiratory	Abnormal results of function study of pulmonary system	31	2.79E-10
QSER1	respiratory	Acute tonsillitis	20	3.35E-10
PRR12	respiratory	Tracheostomy complications	49	6.59E-09
LCP2	respiratory	Chronic obstructive asthma with exacerbation	29	3.37E-08

FRYL	respiratory	Disorders of diaphragm	64	6.10E-08
PCLO	respiratory	Respiratory complications	47	2.79E-07
WDR1	sense organs	Corneal opacity	75	3.00E-10
KMT2A	sense organs	Mastoiditis & related conditions	29	7.27E-10
KIAA1429	sense organs	Pain, swelling or discharge of eye	68	1.71E-09
SPTBN2	sense organs	Tympanosclerosis and middle ear disease related to otitis media	26	1.73E-09
CLASP2	sense organs	Toxic maculopathy of retina	30	2.51E-08
COL1A1	sense organs	Aphakia and other disorders of lens	54	9.89E-08
CACNA1D	sense organs	Corneal edema	47	1.90E-07
CPEB1	sense organs	Otorrhea	72	3.42E-07
TCF20	dermatologic	Other specified erythematous conditions	76	9.76E-10
MYRF	dermatologic	Sarcoidosis	206	1.10E-07
CUX1	genitourinary	Other inflammatory disorders of male genital organs	28	3.44E-09
CSMD3	genitourinary	Acute cystitis	45	3.87E-09
HIVEP1	genitourinary	Acute glomerulonephritis, NOS	28	1.70E-08
GCN1L1	genitourinary	Renal colic	21	2.76E-08
SON	genitourinary	Other inflammatory disorders of male genital organs	28	3.22E-08
ADCY1	genitourinary	Irregular menstrual bleeding	27	4.60E-08
DLGAP1	genitourinary	Renal sclerosis, NOS	72	1.02E-07
MAML2	genitourinary	Urethritis and urethral syndrome	22	1.12E-07
CAPZA1	genitourinary	Stricture/obstruction of ureter	205	1.65E-07
CBFA2T2	genitourinary	Bladder neck obstruction	75	2.44E-07
CASKIN1	genitourinary	Non-proliferative glomerulonephritis	69	3.20E-07
BRWD1	hematopoietic	Hemolytic-uremic syndrome	21	1.23E-08
PHLPP1	hematopoietic	Disorders of iron metabolism	50	3.87E-08
PAN2	hematopoietic	Polycythemia vera, secondary	31	1.09E-07
ATRNL1	hematopoietic	Other hereditary hemolytic anemias	40	1.36E-07
SEC24C	hematopoietic	Defibrination syndrome	38	2.59E-07

Table 2: Significant PheWAS associations with LoF-intolerant genes. A false discovery rate of 1% was enforced to account for multiple testing. A total of 129 significant gene-phenotype associations were identified. PheWAS categories are sorted by the most significant association in that category; gene-phenotype associations are sorted within each category by p-value.

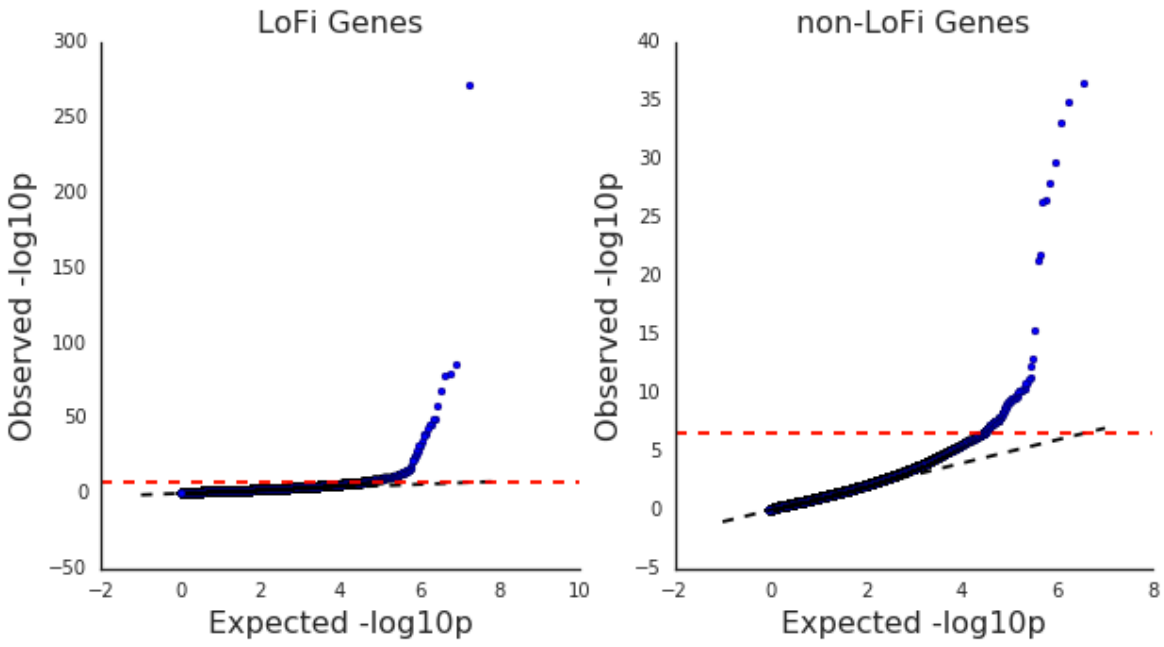

Figure 17: QQ-plot for (a) LoFi and (b) non-LoFi gene-phenotype associations. We observed no p-value inflation, and identified (A) 129 significant associations with LoFi genes and (B) 674 significant associations with non-LoFi genes at a false discovery rate of 1%. The extreme LoFi outlier is the association of rare variants in alpha-1 antitrypsin with alpha-1 antitrypsin deficiency.

Enrichment for clinical and mammalian phenotypes amongst LoFi genes

To determine whether LoFi genes were associated with a clinically distinct collection of phenotypes, we measured whether certain phenotype categories were enriched for significant associations with LoFi genes, relative to non-significant associations, as described in the methods for this chapter. We identified significant depletion for circulatory system phenotypes, and significant enrichment for pregnancy complications and congenital anomalies (Figure 18A). However, we observed these same enrichments amongst non-LoFi genes (Figure 18C), and we observed no significant enrichment or depletion for phenotypes between LoFi and non-LoFi genes with significant associations (p=0.61 to 0.79, Fisher Exact). These shared enrichments may therefore represent differences in the heritability of different phenotype categories or biases in the ascertainment of heritable diseases within a clinical setting. Significant depletion may also identify phenotype categories in which genetic disruption causes severe dysregulation more often resulting embryonic lethality, rather than a disease phenotype.

We next evaluated whether mammalian phenotypes known for the significantly associated genes could provide additional information about disease etiology. We tested for enrichment or depletion for mammalian phenotypes derived from mouse model systems for all LoFi and non-LoFi genes significantly associated with any clinical phenotype. Despite the consistency in phenotypic associations observed, LoFi and non-LoFi genes displayed distinct patterns of mammalian phenotype enrichment (Figures 18B and 18D). Notably, significantly associated LoFi and non-LoFi genes were significantly enriched and depleted for phenotypes relating to embryonic development in mice, respectively. Despite their

association with similar phenotype categories, this may indicate that LoFi genes are active earlier in development, increasing the severity of any deleterious effects. Alternatively, these differences may suggest that although there is no significant difference between the clinical phenotypes associated with LoFi and non-LoFi genes, the biological mechanisms underlying their contribution to disease are different.

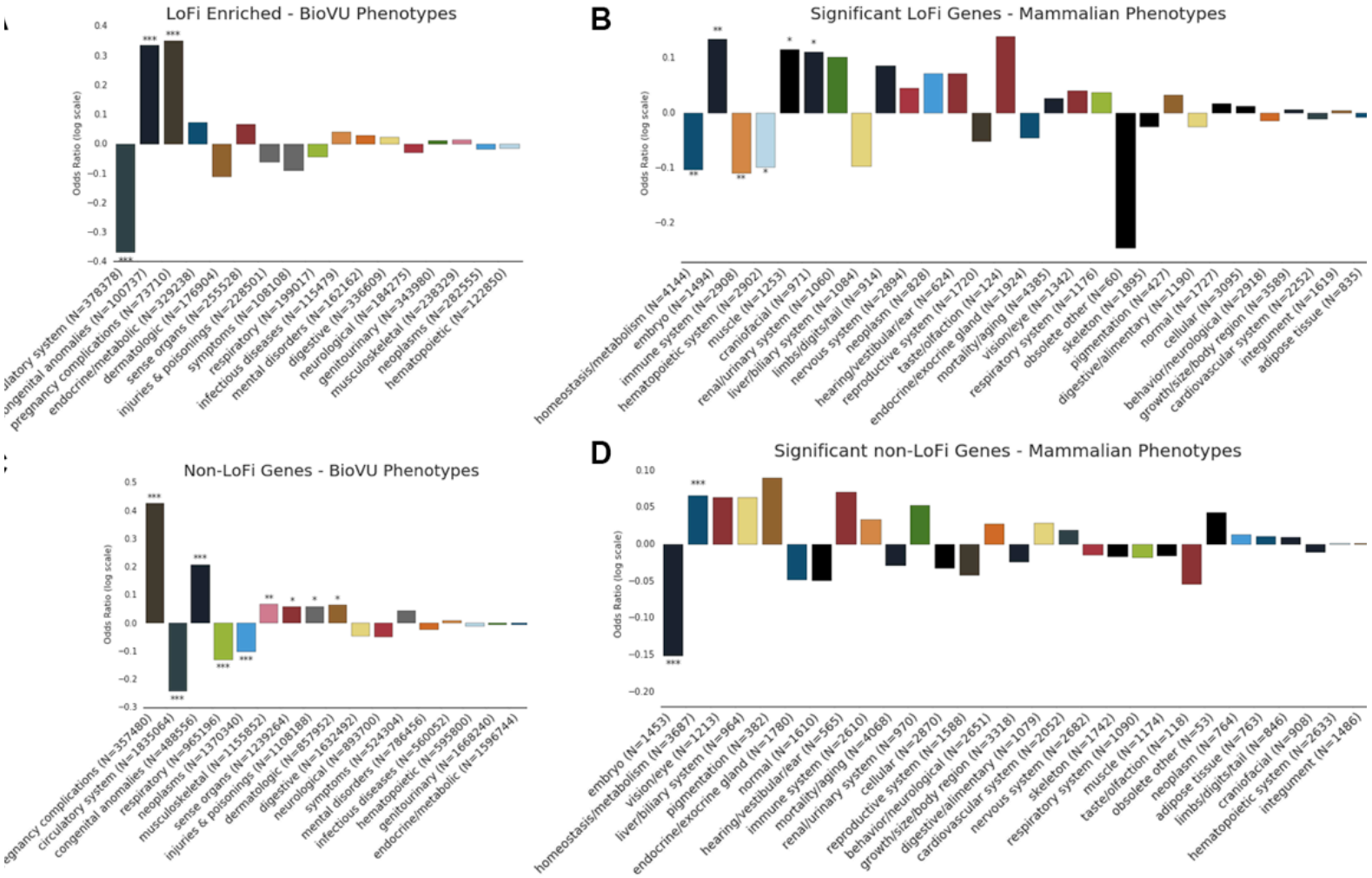

Figure 18: Significant associations with LoFi genes are not evenly distributed among phenotype categories. (A) LoFi genes are significantly depleted for associations with circulatory system and respiratory phenotypes, and significantly enriched for associations with congenital anomalies, pregnancy complications, injuries and poisonings, and sense organ phenotypes. (B) Non-LoFi genes that significantly associated with any clinical phenotype are significantly depleted for mammalian phenotypes relating to homeostasis and metabolism, the immune system, and the hematopoietic system, and significantly enriched embryonic, muscle, and craniofacial phenotypes.

Biological validation using existing mouse model phenotypes

Approximately 72% of LoFi genes are not currently associated with a human disease phenotype[7]. The discrepancy between the apparent functional importance of LoFi genes and their lack of known disease associations makes them an

interesting target for PheWAS. However, this property also limits the amount of phenotypic and functional support for any significant associations, and could reduce the likelihood of these associations being carried forward for future research. To estimate the proportion of statistically significant gene-phenotype associations with existing support from model systems, we incorporated data from the Mouse Genome Informatics (MGI) Human Gene Disease Connection (HGDC). Of the 129 significant LoFi gene-phenotype associations, 26 (20%) were consistent with observed mammalian phenotype categories (Table 3), twice the proportion of significant non-LoFi gene-phenotype associations with MGI support (69 of 674, ~10%). For example, PRDM2 is significantly associated with benign neoplasms of the eye ($p=3.14x10^{-07}$), and is reported by MGI to increase the incidence of tumors in mice. Furthermore, PRDM2 is included in the COSMIC Cancer Gene Consensus as a tumor suppressor gene, and contains somatic mutations found in human cancers[18]. However, we found no previous mention of germline variation in PRDM2 playing a role in neoplasms. While specific phenotypes observed in mice rarely match the specific clinical phenotypes associated with each gene, this data supports the biological plausibility of significant associations.

Gene	p-value	PheWAS Category	PheWAS Description	MGI Phenotype Category
SLC4A4	6.35E-22	endocrine/metabolic	Alpha-1-antitrypsin deficiency	homeostasis/metabolism
DST	2.01E-11	endocrine/metabolic	Lipoprotein disorders	homeostasis/metabolism
CD72	3.84E-10	endocrine/metabolic	Pituitary hyperfunction	homeostasis/metabolism
IGF2R	1.23E-08	endocrine/metabolic	Other immunological findings	endocrine/exocrine gland
USP47	3.09E-08	endocrine/metabolic	Autoimmune disease NEC	homeostasis/metabolism
HTT	5.05E-13	infectious diseases	Sexually transmitted infections (not HIV or hepatitis)	immune system
DNMT3B	6.44E-09	infectious diseases	H. pylori	immune system
PAX3	6.42E-08	infectious diseases	Sexually transmitted infections (not HIV or hepatitis)	immune system
PRDM1	1.53E-07	infectious diseases	Sexually transmitted infections (not HIV or hepatitis)	immune system
PIKFYVE	5.44E-11	congenital anomalies	Other congenital anomalies of lower limb, including pelvic girdle	growth/size/body region
NUP85	1.42E-08	congenital anomalies	Obstructive genitourinary defect	mortality/aging
WDR1	3.00E-10	sense organs	Corneal opacity	hearing/vestibular/ear
KMT2A	7.27E-10	sense organs	Mastoiditis & related conditions	hearing/vestibular/ear
COL1A1	9.89E-08	sense organs	Aphakia and other disorders of lens	hearing/vestibular/ear
CACNA1D	1.90E-07	sense organs	Corneal edema	hearing/vestibular/ear
ERBB2	6.01E-10	circulatory system	Polyarteritis nodosa	cardiovascular system
BMP7	1.81E-09	mental disorders	Acute reaction to stress	behavior/neurological
CELSR3	2.12E-08	mental disorders	Mental disorders durring/after pregnancy	behavior/neurological
SHANK2	5.67E-08	mental disorders	Vascular dementia	behavior/neurological
CUX1	3.44E-09	Genitourinary	Other inflammatory disorders of male genital organs	renal/urinary system
CASKIN1	3.20E-07	Genitourinary	Non-proliferative glomerulonephritis	renal/urinary system
LCP2	3.37E-08	Respiratory	Chronic obstructive asthma with exacerbation	respiratory system
RELN	4.40E-08	pregnancy complications	Miscarriage; stillbirth	reproductive system

PRKCQ	1.38E-07	Neoplasms	Radiotherapy	neoplasm
PTPRT	2.58E-07	Neoplasms	Bone marrow or stem cell transplant	neoplasm
PRDM2	3.14E-07	Neoplasms	Benign neoplasm of eye, uveal	neoplasm

Table 3: Significant LoFi gene-phenotype associations with mouse model support. Of the 129 LoFi genes significantly associated with BioVU phenotypes, 26 match a mammalian phenotype category reported by Mouse Genome Informatics. Each PheWAS category was assigned to the closest semantically-matching MGI phenotype category.

4.4 Conclusion

The extreme depletion of protein-truncating variants within a gene is evidence of selective constraint against gene loss-of-function and suggests that the loss of gene function leads to embryonic lethality. Despite clear evidence for their functional importance, disease associations are known for less than a third of genes predicted to be loss-of-function intolerant. This study is the first comprehensive, gene-level PheWAS characterizing the phenotypic impact of missense variation within loss-of-function intolerant genes. We present 129 novel associations between LoFi genes and clinical phenotypes, 26 of which (20%) are concordant with existing mouse phenotypes. We also demonstrate that the phenotypic spectrum of LoFi genes is not statistically distinct from other genes, suggesting that the importance of the gene and/or severity of the phenotype, rather than the class of phenotype, are what drive the selective constraint against LoF variants.

We identified no significant enrichment or depletion for particular clinical phenotypes among significantly associated LoFi genes, relative to non-LoFi genes. We interpret this to mean that these genes are not responsible for a clinically distinct subset of phenotype, but rather that these genes are critical for the function of many physiological systems. The shared enrichments and depletions we observe likely reflect the statistical power available for different phenotypes within our clinical cohort. Alternatively, these statistics may be suggestive of both the genetic heritability and severity of perturbations to different physiological systems. For example, the significant depletion for circulatory system phenotypes may indicate low genetic heritability, or it may indicate that variants disrupting circulatory system phenotypes are significantly more likely to result in embryonic lethality, and are thus less likely to be observed in our clinical cohort. This interpretation is supported by research in mouse model systems, where the mutations most likely to disrupt early development and result in early lethality are those causing cardiovascular defects. Within this logical framework, it may be possible to infer genetic association from significant phenotypic depletion; despite a significant depletion for significant associations, cardiovascular phenotypes may thus be the most common cause of embryonic lethality in LoFi gene knockouts.

We identified significant associations with human phenotypes for 129 LoFi genes, 89 of which had no previous association with human disease; existing mammalian phenotypes from mouse model systems support many of these associations. The selective constraint against protein loss-of-function suggests an extreme relationship between these genes and their phenotypes. Combined with the mouse-model support, pharmaceutical interventions targeting these genes may identify novel therapies relating to the associated traits. For example, *CELSR3*

encodes CELR3, a G-coupled protein receptor (GPCR), and is significantly associated with mental disorders during or after pregnancy. CELR3 is also involved in dopaminergic and serotonergic neuron axon guidance during development[94], and is annotated to interact with the pregnancy-specific beta-1-glycoprotein (PSG) family of proteins[88]. It may be that the association of *CELSR3* with pregnancy-related mental disorders is related to interactions between CELR3 and PSG proteins during pregnancy. Although additional functional characterization is clearly necessary, GPCRs are highly druggable targets, and inhibition of CELR3-PSB interactions during pregnancy may provide a novel treatment for pregnancy-related disorders like postpartum depression.

This study is limited by the use of an exome genotyping array, rather than whole-exome sequencing. We are thus limited to known variants, which are already at very low frequency within LoFi genes. This limitation is somewhat mitigated by the inclusion of other protein-coding variation, but the analysis would be best conducted in a large whole-exome sequencing dataset linked with electronic medical records[111]. Another limitation of the study is the use of samples exclusively of European ancestry. While this decision is intended to reduce confounding within the statistical analysis, it may also limit the generalizability of our results to individuals of non-European descent.

We conducted this study under an assumption of allelic heterogeneity, and chose to perform our gene-level association tests using SKAT-O. We note that genes with significant associations to any phenotype had approximately twice as many variants as all genes (median 14 and 7 variants per gene, respectively). Using single-marker association tests to follow-up significant gene-phenotype associations, we estimate that significant associations are typically driven by ~14% of variants in each gene, with a median of 2 nominally significant ($p<0.05$) single-marker associations per significantly associated gene. It is possible that with whole-exome sequencing, the proportion of nominally significant single-marker associations contributing to significant gene-level associations would be increased. However, caution should be used when interpreting the current results within the framework of allelic heterogeneity.

Chapter 5

Discussion

In this dissertation, we explore ways in which constraint on rare protein-coding variation can be used to better understand the genetic basis for human disease. Specifically, we focus on two structure-based methods, one for variant pathogenicity prediction and another for the detection of constraints on the distribution of somatic mutations, and one gene-level method for phenotypic discovery through aggregate analysis of rare protein-coding variation in loss-of-function intolerant genes.

Our first analysis explored the use of existing phenotypic associations, combined with protein structural information, to evaluate a method for variant pathogenicity prediction based on spatial proximity to known pathogenic variants. We focused our analysis on disease-causing variants in *RTEL1,* with a particular interest in predicting variants that increase risk for FIP. This analysis served as a demonstration of the practical application of spatial predictors for classifying variants of unknown significance. The methodology is dependent only on the availability of protein structural information (whether experimentally derived or computationally predicted) and the assumption that disease-causing variants are spatially clustered within the protein structure. Solved proteins structures are available for nearly a quarter of human proteins, and the inclusion of computationally predicted homology models can increase that coverage to over three quarters of human proteins. The tendency for cancer-associated somatic mutations to form spatial clusters in protein sequence and structure is well established[26], and evidence for spatial clustering has likewise been observed for germline disease-causing variants[23,101] and demonstrated in our previous work. Furthermore, our previous research demonstrated that spatial proximity to pathogenic variation was a useful predictor of variant pathogenicity in hundreds of proteins. Thus, the methodology proposed here will likely be broadly useful in the identification of disease regions of interest within protein structure and variant pathogenicity prediction. However, predictors relying exclusively on spatial statistics cannot discern between variants affecting the same amino acid position (which may have drastically different severities), and lack many other informative features that can be derived from protein structures. Therefore, spatial statistics will be most effective when integrated with other pathogenicity prediction methods that take this information into account.

Our results demonstrate that considering the 3D spatial landscape of missense variation in RTEL1 has the potential to improve pathogenicity prediction and identify functional regions of protein structure important to the development of disease. We implicate the ATP-binding cleft between helicase domains I and II as well as the DNA-binding pore along helicase domain II as functional regions of RTEL1 contributing to the development of FIP. The similar distributions of disease-associated variants and a significant correlation with ATPase activity in the homologous protein XPD support this finding and suggest that including additional variants from homologous proteins may improve predictive power and

discover shared biochemical etiology. More generally, we propose incorporating the spatial distributions of known pathogenic and neutral variation into pathogenicity prediction methods to complement existing predictive features, particularly for proteins in which pathogenic variants appear to form clusters within protein structure. Ultimately, the use of this information has the potential to enhance the utility of genetic data in elucidating the etiology of FIP and other heritable diseases.

The measure of pathogenic proximity we developed to classify missense variants of unknown significance in the protein structure of RTEL1 is built upon Euclidean distance and proximity comparisons. This predictor ignores a wealth of useful information about population genetics and mutagenic processes, like variant allele frequencies, inconsistent mutation rates across a gene, and differences in the number and severity of amino acid substitutions possible for any given single-nucleotide polymorphism. To fully exploit the predictive potential of spatial information, we need to develop a robust algorithm with a solid foundation in selective constraint. Thoughtful approaches are currently in development for sequence-based analyses at both the whole-gene[29,30] and sub-gene[8,112] levels. We suggest a modular solution for measuring selective constraint within protein structure that incorporates structural information into a well-conceived, sequence-based metric. This approach would focus our efforts on the integration of structural information, and facilitate iterative improvements to, or replacement of the underlying constraint metrics developed by other researchers. This approach would best allocate the expertise of those involved and produce a well-supported, highly maintainable measure of selective constraint in protein structure.

In our analysis of somatic mutations, we used a consistent statistical framework to identify significant clustering in solved and predicted protein structures. Structural analysis of these spatial clusters has the potential to uncover previously unknown disease etiology and suggest potential drug targets. More broadly, our results indicate that selective constraint within the tumor context influences the spatial distribution of somatic mutations in protein structure, and support the use of large reference datasets to highlight regions of tumorigenic importance. In contrast to our analysis of known disease-causing variation in the protein structure of RTEL1, this analysis demonstrates how the spatial distribution of unlabeled protein-coding variants in protein structure can itself be used to identify functional subsets of variants and potentially identify driver mutations. However, compared to our previous analysis of germline disease-causing variants, we find that somatic mutations are much less likely to be clustered in protein structure. This may be partially attributable to an abundance of passenger mutations disrupting the detection of driver mutation clusters, or may accurately detect relaxed constraint in the somatic or tumor context. Regardless of the cause, our results indicate that the spatial analysis of germline disease-causing variation will produce as many, and likely more insights about the genetic basis of inherited disease, relative to the study of somatic driver mutations in cancer. This conclusion further supports development of the structure-based measure of selective constraint described above.

Not all genes contain well-characterized variants with known associations to disease. Loss-of-function intolerant (LoFi) genes are under extreme selective

constraint, and are likely of critical functional importance to human health and viability. Despite evidence of their functional importance, the phenotypic relevance of these genes is largely unknown. We leveraged dense whole-exome genotyping of protein-coding variation in an EMR-linked biobank to quantify the phenome-wide impact of LoF-intolerant genes, capturing over one hundred significant associations with clinical phenotypes, many of which constitute the first association with human disease.

In future work, we plan to replicate these significant associations using whole-exome sequencing data linked with electronic medical records, which will greatly improve the coverage of low-frequency and loss-of-function variants, and enable the detection of *de novo* and ultra-rare variants. We propose to carry out the replication analysis using data from the Geisinger-Regeneron DiscovEHR cohort[111], which is composed primarily (98%) of individuals of European ancestry, which are likely to match the genetic background of our study population; the DiscovEHR cohort also more than doubles our discovery sample size. We will first attempt to replicate our findings using only the genetic variants included in our discovery analysis, and then conduct a follow-up study in which all available whole-exome sequencing data is considered. However, there are still potential confounding factors inherent with replicating our results in another clinical cohort. Most notably, while our study populations are well-matched on genetic ancestry, they are ascertained from geographically distinct locations (Tennessee/Kentucky and Pennsylvania), which introduces an unknown number of environmental differences that may affect the replication of any associations contingent on environmental exposures. Additionally, because the clinical phenotypes used in our PheWAS are derived from insurance billing codes, different hospitals (especially those in different states) are likely to have different prescribing and billing practices. For example, if related billing codes are reimbursed differently in different states or by different insurance companies, two hospitals may assign different billing codes for the same disease to maximize reimbursement. Some of this variance in billing will be mitigated by the aggregation of related ICD9 codes into PheWAS codes, but it is important to recognize that the replication of a PheWAS association may not necessarily require an exact match with the original PheWAS code. For phenotypes exactly or approximately replicating in the DiscovEHR cohort, we can then pursue more sophisticated phenotyping algorithms that incorporate additional information (e.g. procedure codes, quantitative lab measurements, and clinical notes) from the EMR to more accurately identify cases and controls from the clinical cohort.

Ultimately, phenotypic associations with LoF-intolerant genes are expected to reveal genes critical to early development and may also assist in the identification of highly effective drug targets for therapeutic intervention. Our analysis takes the first steps in identifying gene-level associations with clinical phenotypes; they also differ from previous genome-wide associations with protein-truncating variants (PTVs) in LoF-intolerant genes identified by Ganna *et al*. This unpublished study predominantly identified genome-wide associations with psychiatric phenotypes, while we identified gene-level associations no significant enrichment for psychiatric disorders. Ganna *et al*. also found that their phenotypic associations were specific to PTVs in LoF-intolerant genes, while we found that

our phenotype enrichments were not specific to LoF-intolerant genes, and reflected a similar pattern of phenotype enrichment and depletion as other genes. These differences may be attributable to sample ascertainment strategies; samples from Ganna *et al.* were ascertained specifically for psychiatric disorders, while our samples are derived from a clinical cohort with no phenotypic ascertainment strategy. However, the same ascertainment strategies used for the psychiatric cohorts were also used for other phenotypic cohorts that showed no significant associations. The disagreement may instead be attributable to differences in whole-exome sequencing and genotyping; however, it is unclear whether this difference is driven by access to lower frequency variants (low-frequency vs. ultra-rare and *de novo*) or by the types of variants available for analysis (all protein-coding vs. loss-of-function). In either case, if differences between our results are attributable to whole-exome sequencing and whole-exome genotyping, we should observe increased similarity between the results of our replication analysis and those reported by Ganna *et al*. If we do not, it is more likely that carefully defined cohorts are required to identify significant associations with psychiatric disorders. Finally, to control for differences between genome-wide and gene-level analyses, we will perform an analogous genome-wide aggregate association test for direct comparison with the methodology employed by Ganna *et al.* As the only two studies characterizing the phenotypic impact of LoF-intolerant genes, it is important that any difference in the phenotypic associations identified is carefully addressed and well understood.

This dissertation makes a significant advance in our understanding of how selective constraint on protein-coding genetic variants can help to explain their contribution to disease. We demonstrate utility in variant pathogenicity prediction, the detection of putative driver mutations in cancer, and the identification of novel phenotype associations with highly constrained genes. These approaches will continue to improve as whole-exome and whole-genome sequencing becomes increasingly prevalent, and developing these methods now is critical to maximizing the scientific impact of this data.

BIBLIOGRAPHY

1. Bustamante, C.D. (2005). Natural selection on protein-coding genes in the human genome. Nature *437*, 1153–1157.
2. Consortium, T. 1000 genomes project, Durbin, R.M., Altshuler, D.L., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Collins, F.S., De La Vega, F.M., Donnelly, P., et al. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.
3. Boyko, A.R., Williamson, S.H., Indap, A.R., Degenhardt, J.D., Hernandez, R.D., Lohmueller, K.E., Adams, M.D., Schmidt, S., Sninsky, J.J., Sunyaev, S.R., et al. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. *4*,.
4. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.
5. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., et al. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216–220.
6. Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C.R., Lim, E.P., Kalyanaraman, N., et al. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. *22*, 231–238.
7. Lek, M., Karczewski, K.J., Minikel, E. V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
8. Samocha, K.E. (2014). A framework for the interpretation of de novo mutation in human disease. Nat. Genet. *46*, 944–950.
9. Berman, H.M. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.
10. Pieper, U., Webb, B.M., Barkan, D.T., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., Yang, Z., Meng, E.C., Pettersen, E.F., Huang, C.C., et al. (2011). ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res. *39*, D465-74.
11. Dixon, P.M. (2002). Ripley's K function. Encycl. Environmetrics *3*, 1796–1803.
12. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. *44*, D862–D868.
13. Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814.
14. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
15. Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. Hum. Mutat. *32*, 358–368.
16. Hu, H., Huff, C.D., Moore, B., Flygare, S., Reese, M.G., and Yandell, M. (2013). VAAST 2.0: Improved Variant Classification and Disease-Gene Identification Using a Conservation-Controlled Amino Acid Substitution Matrix. Genet. Epidemiol. *37*, 622–634.
17. Baugh, E.H., Simmons-Edler, R., Mueller, C.L., Alford, R.F., Volfovsky, N., Lash, A.,

and Bonneau, R. (2015). Robust Classification of Protein Variation Using Structural Modeling and Large-Scale Data Integration. Preprint *XX*, 1–6.
18. Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., et al. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. *43*, D805-11.
19. Chang, K., Creighton, C.J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., Ally, A., Balasundaram, M., Birol, I., Butterfield, Y.S.N., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120.
20. Stehr, H., Jang, S.-H.J., Duarte, J.M., Wierling, C., Lehrach, H., Lappe, M., and Lange, B.M.H. (2011). The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. Mol. Cancer *10*, 54.
21. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
22. Kamburov, A., Lawrence, M.S., Polak, P., Leshchiner, I., Lage, K., Golub, T.R., and Wheeler, D. (2015). Comprehensive assessment of cancer missense mutation clustering in protein structures. 1–10.
23. Meyer, M.J., Lapcevic, R., Romero, A.E., Yoon, M., Das, J., Beltrán, J.F., Mort, M., Stenson, P.D., Cooper, D.N., Paccanaro, A., et al. (2016). Mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. Hum. Mutat. n/a-n/a.
24. Collin Tokheim, Rohit Bhattacharya, Noushin Niknafs, Derek M Gygax, Rick Kim, M., and Ryan, David Masica, R.K. (2016). Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. Cancer Res.
25. Niu, B., Scott, A.D., Sengupta, S., Bailey, M.H., Batra, P., Ning, J., Wyczalkowski, M.A., Liang, W.-W., Zhang, Q., McLellan, M.D., et al. (2016). Protein-structure-guided discovery of functional mutations across 19 cancer types. Nat. Genet.
26. Porta-Pardo, E., Kamburov, A., Tamborero, D., Pons, T., Grases, D., Valencia, A., Lopez-Bigas, N., Getz, G., and Godzik, A. (2017). Comparison of algorithms for the detection of cancer drivers at subgene resolution. Nat. Methods *14*,.
27. Araya, C.L., Cenik, C., Reuter, J.A., Kiss, G., Pande, V.S., Snyder, M.P., and Greenleaf, W.J. (2016). Systematic identification of significantly mutated regions reveals a rich landscape of functional molecular alterations across cancer genomes. Nat. Genet. 20875.
28. Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2013). OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics *29*, 2238–2244.
29. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S., and Goldstein, D.B. (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLoS Genet. *9*,.
30. Lek, M. (2015). Analysis of protein-coding genetic variation in 60,706 humans. 1–26.
31. Ganna, A., Satterstrom, K., Zekavat, S., Das, I., Kurki, M., Churchhouse, C., Alfoldi, J., Martin, A., Havulinna, A., Byrnes, A., et al. (2017). Quantifying the impact of rare and ultra-rare coding variation across the phenotypic spectrum. bioRxiv.
32. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.

33. Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Christiani, D.C., Wurfel, M.M., and Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. *91*, 224–237.
34. Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D.R., Roden, D.M., and Crawford, D.C. (2010). PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics *26*, 1205–1210.
35. Bush, W.S., Oetjens, M.T., and Crawford, D.C. (2016). Unravelling the human genome-phenome relationship using phenome-wide association studies. Nat. Rev. Genet. *17*, 129–145.
36. Pendergrass, S.A., Brown-Gentry, K., Dudek, S., Frase, A., Torstenson, E.S., Goodloe, R., Ambite, J.L., Avery, C.L., Buyske, S., Bůžková, P., et al. (2013). Phenome-Wide Association Study (PheWAS) for Detection of Pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. PLoS Genet. *9*,.
37. Namjou, B., Marsolo, K., Caroll, R.J., Denny, J.C., Ritchie, M.D., Verma, S.S., Lingren, T., Porollo, A., Cobb, B.L., Perry, C., et al. (2014). Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links PLCL1 to speech language development and IL5-IL13 to Eosinophilic Esophagitis. Front. Genet. *5*,.
38. Simonti, C.N., Vernot, B., Bastarache, L., Bottinger, E., Carrell, D.S., Chisholm, R.L., Crosslin, D.R., Hebbring, S.J., Jarvik, G.P., Kullo, I.J., et al. (2016). The phenotypic legacy of admixture between modern humans and Neandertals. Science (80-.). *351*, 737–741.
39. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.
40. Hemnes, A.R., Zhao, M., West, J., Newman, J.H., Rich, S., Archer, S.L., Robbins, I.M., Blackwell, T.S., Cogan, J., Loyd, J.E., et al. (2016). Critical Genomic Networks and Vasoreactive Variants in Idiopathic Pulmonary Arterial Hypertension. Am. J. Respir. Crit. Care Med.
41. de Jesus Perez, V.A., Yuan, K., Lyuksyutova, M.A., Dewey, F., Orcholski, M.E., Shuffle, E.M., Mathur, M., Yancy, L., Rojas, V., Li, C.G., et al. (2014). Whole-exome sequencing reveals TopBP1 as a novel gene in idiopathic pulmonary arterial hypertension. Am. J. Respir. Crit. Care Med. *189*, 1260–1272.
42. Eyries, M., Montani, D., Girerd, B., Perret, C., Leroy, A., Lonjou, C., Chelghoum, N., Coulet, F., Bonnet, D., Dorfmüller, P., et al. (2014). EIF2AK4 mutations cause pulmonary veno-occlusive disease, a recessive form of pulmonary hypertension. Nat. Genet. *46*, 65–69.
43. Ma, L., Roman-Campos, D., Austin, E.D., Eyries, M., Sampson, K.S., Soubrier, F., Germain, M., Trégouët, D.-A., Borczuk, A., Rosenzweig, E.B., et al. (2013). A novel channelopathy in pulmonary arterial hypertension. N. Engl. J. Med. *369*, 351–361.
44. Austin, E.D., Ma, L., LeDuc, C., Berman Rosenzweig, E., Borczuk, A., Phillips, J.A., Palomero, T., Sumazin, P., Kim, H.R., Talati, M.H., et al. (2012). Whole exome sequencing to identify a novel gene (caveolin-1) associated with human pulmonary arterial hypertension. Circ. Cardiovasc. Genet. *5*, 336–343.
45. Cogan, J.D., Kropski, J. a., Zhao, M., Mitchell, D.B., Rives, L., Markin, C., Garnett, E.T., Montgomery, K.H., Mason, W.R., McKean, D.F., et al. (2015). Rare Variants in RTEL1 Are Associated with Familial Interstitial Pneumonia. Am. J. Respir. Crit. Care

Med. *191*, 646–655.
46. Stuart, B.D., Choi, J., Zaidi, S., Xing, C., Holohan, B., Chen, R., Choi, M., Dharwadkar, P., Torres, F., Girod, C.E., et al. (2015). Exome sequencing links mutations in PARN and RTEL1 with familial pulmonary fibrosis and telomere shortening. Nat. Genet. *47*, 512–517.
47. Caroline Kannengiesser1, 2*, Raphael Borie3*, Christelle Ménard1, Marion Réocreux1, P., Nitschké2, 4, Steven Gazal2, 5, 6, Hervé Mal7, Jacques Cadranel8, 9, Hilario Nunes10, 11, D., Valeyre10, 11, Jean François Cordier, 13, Isabelle Callebaut14, Catherine Boileau1, 2, V., and Cottin12, 13, Bernard Grandchamp1, 2, Patrick Revy15, Bruno Crestani 2, 3 (2015). Heterozygous RTEL1 mutations is a major cause of familial pulmonary fibrosis. Eur. Respir. J.
48. Diaz de Leon, A., Cronkhite, J.T., Katzenstein, A.L.A., Godwin, J.D., Raghu, G., Glazer, C.S., Rosenblatt, R.L., Girod, C.E., Garrity, E.R., Xing, C., et al. (2010). Telomere lengths, pulmonary fibrosis and telomerase (TERT) Mutations. PLoS One *5*,.
49. Cronkhite, J.T., Xing, C., Raghu, G., Chin, K.M., Torres, F., Rosenblatt, R.L., and Garcia, C.K. (2008). Telomere shortening in familial and sporadic pulmonary fibrosis. Am. J. Respir. Crit. Care Med. *178*, 729–737.
50. Armanios, M., Alder, J.K., Chen, J.J.-L., Lancaster, L., Danoff, S., Su, S., Cogan, J.D., Vulto, I., Xie, M., Qi, X., et al. (2008). Short telomeres are a risk factor for idiopathic pulmonary fibrosis. Proc. Natl. Acad. Sci. U. S. A. *105*, 13051–13056.
51. Alder, J.K., Barkauskas, C.E., Limjunyawong, N., Stanley, S.E., Kembou, F., Tuder, R.M., Hogan, B.L.M., Mitzner, W., and Armanios, M. (2015). Telomere dysfunction causes alveolar stem cell failure. Proc. Natl. Acad. Sci. *112*, 201504780.
52. Povedano, J.M., Martinez, P., Flores, J.M., Mulero, F., and Blasco, M. a (2015). Mice with Pulmonary Fibrosis Driven by Telomere Dysfunction. Cell Rep. *12*, 286–299.
53. Chen, R., Zhang, K., Chen, H., Zhao, X., Wang, J., Li, L., Cong, Y., Ju, Z., Xu, D., Williams, B.R.G., et al. (2015). Telomerase deficiency causes alveolar stem cell senescence-associated low-grade inflammation in lungs. J. Biol. Chem. *290*, 30813–30829.
54. Armanios, M., Chen, J.J.-L., Cogan, J.D., Alder, J.K., Ingersoll, R.G., Markin, C., Lawson, W.E., Xie, M., Vulto, I., Phillips, J. a, et al. (2007). Telomerase mutations in families with idiopathic pulmonary fibrosis. N. Engl. J. Med. *356*, 1317–1326.
55. Tsakiri, K.D., Cronkhite, J.T., Kuan, P.J., Xing, C., Raghu, G., Weissler, J.C., Rosenblatt, R.L., Shay, J.W., and Garcia, C.K. (2007). Adult-onset pulmonary fibrosis caused by mutations in telomerase. Proc. Natl. Acad. Sci. U. S. A. *104*, 7552–7557.
56. Walne, A.J., Vulliamy, T., Kirwan, M., Plagnol, V., and Dokal, I. (2013). Constitutional mutations in RTEL1 cause severe dyskeratosis congenita. Am. J. Hum. Genet. *92*, 448–453.
57. Deng, Z., Glousker, G., Molczan, A., Fox, A.J., Lamm, N., Dheekollu, J., Weizman, O.-E., Schertzer, M., Wang, Z., Vladimirova, O., et al. (2013). Inherited mutations in the helicase RTEL1 cause telomere dysfunction and Hoyeraal-Hreidarsson syndrome. Proc. Natl. Acad. Sci. U. S. A. *110*, E3408-16.
58. Ballew, B.J., Joseph, V., De, S., Sarek, G., Vannier, J.B., Stracker, T., Schrader, K. a., Small, T.N., O'Reilly, R., Manschreck, C., et al. (2013). A Recessive Founder Mutation in Regulator of Telomere Elongation Helicase 1, RTEL1, Underlies Severe Immunodeficiency and Features of Hoyeraal Hreidarsson Syndrome. PLoS Genet. *9*,.
59. Ballew, B.J., Yeager, M., Jacobs, K., Giri, N., Boland, J., Burdett, L., Alter, B.P., and Savage, S. a. (2013). Germline mutations of regulator of telomere elongation helicase 1,

RTEL1, in Dyskeratosis congenita. Hum. Genet. *132*, 473–480.
60. Hanna, S., Béziat, V., Jouanguy, E., Casanova, J.L., and Etzioni, A. (2015). A homozygous mutation of RTEL1 in a child presenting with an apparently isolated natural killer cell deficiency. J. Allergy Clin. Immunol. *136*, 1113–1114.
61. Moriya, K., Niizuma, H., Rikiishi, T., Yamaguchi, H., Sasahara, Y., and Kure, S. (2016). Novel Compound Heterozygous RTEL1 Gene Mutations in a Patient With Hoyeraal-Hreidarsson Syndrome. Pediatr. Blood Cancer *63*, 1683–1684.
62. Le Guen, T., Jullien, L., Touzot, F., Schertzer, M., Gaillard, L., Perderiset, M., Carpentier, W., Nitschke, P., Picard, C., Couillault, G., et al. (2013). Human RTEL1 deficiency causes Hoyeraal-Hreidarsson syndrome with short telomeres and genome instability. Hum. Mol. Genet. *22*, 3239–3249.
63. Travis, W.D., Costabel, U., Hansell, D.M., King, T.E., Lynch, D.A., Nicholson, A.G., Ryerson, C.J., Ryu, J.H., Selman, M., Wells, A.U., et al. (2013). An official American Thoracic Society/European Respiratory Society statement: Update of the international multidisciplinary classification of the idiopathic interstitial pneumonias. Am. J. Respir. Crit. Care Med. *188*, 733–748.
64. Durham, E., Dorr, B., Woetzel, N., Staritzbichler, R., and Meiler, J. (2009). Solvent accessible surface area approximations for rapid and accurate protein structure prediction. J. Mol. Model. *15*, 1093–1108.
65. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning (Springer).
66. Goldenberg, O., Erez, E., Nimrod, G., and Ben-Tal, N. (2009). The ConSurf-DB: Pre-calculated evolutionary conservation profiles of protein structures. Nucleic Acids Res. *37*, 323–327.
67. McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics *26*, 2069–2070.
68. Kurowski, M.A., and Bujnicki, J.M. (2003). GeneSilico protein structure prediction meta-server. Nucleic Acids Res. *31*, 3305–3307.
69. Söding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. *33*, W244-8.
70. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2014). The I-TASSER Suite: protein structure and function prediction. Nat. Methods *12*, 7–8.
71. Fernandez-Fuentes, N., Madrid-Aliste, C.J., Rai, B.K., Fajardo, J.E., and Fiser, A. (2007). M4T: a comparative protein structure modeling server. Nucleic Acids Res. *35*, W363-8.
72. Wallner, B., and Elofsson, A. (2005). Pcons5: combining consensus, structural evaluation and fold recognition scores. Bioinformatics *21*, 4248–4254.
73. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J.E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. *10*, 845–858.
74. Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., and Xu, J. (2012). Template-based protein structure modeling using the RaptorX web server. Nat. Protoc. *7*, 1511–1522.
75. Kim, D.E., Chivian, D., and Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. *32*, W526-31.
76. Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Gallo Cassarino, T., Bertoni, M., Bordoli, L., et al. (2014). SWISS-MODEL: modelling

protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. *42*, W252-8.
77. Kuper, J., Wolski, S.C., Michels, G., and Kisker, C. (2011). Functional and structural studies of the nucleotide excision repair helicase XPD suggest a polarity for DNA translocation. EMBO J. *31*, 494–502.
78. Fan, L., Fuss, J., Cheng, Q., Arvai, A., and Hammel, M. (2008). XPD helicase structures and activities: insights into the cancer and aging phenotypes from XPD mutations. Cell.
79. Kim, K., Oh, J., Han, D., Kim, E., Lee, B., and Kim, Y. (2006). Crystal structure of PilF: functional implication in the type 4 pilus biogenesis in Pseudomonas aeruginosa. Biochem. Biophys. Res.
80. Sawaya, M.R., Chan, S., Han, G.W., and Perry, L.J. (2006). Crystal Structure of a Ten A Homolog/Thi-4 Thiaminase from Pyrobaculum Aerophilum. Protein Data Bank.
81. Coquille, S., Filipovska, A., Chia, T., and Rajappa, L. (2014). An artificial PPR scaffold for programmable RNA recognition. Nat. Commun.
82. Rapley, J., Tybulewicz, V., and Rittinger, K. (2008). Crucial structural role for the PH and C1 domains of the Vav1 exchange factor. EMBO Rep.
83. Vollmar, M., Ayinampudi, V., Cooper, C., Guo, K., Krojer, T., Muniz, J.R.C., von Delft, F., Weigelt, J., Arrowsmith, C.H., Bountra, C., et al. (2012). Crystal structure of the N-terminal domain of human Cul4B at 2.57A resolution. Protein Data Bank.
84. Tyka, M., Keedy, D., André, I., DiMaio, F., and Song, Y. (2011). Alternate states of proteins revealed by detailed energy landscape mapping. J. Mol.
85. Mandell, D., Coutsias, E., and Kortemme, T. (2009). Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat. Methods.
86. Uringa, E.J., Youds, J.L., Lisaingo, K., Lansdorp, P.M., and Boulton, S.J. (2011). RTEL1: An essential helicase for telomere maintenance and the regulation of homologous recombination. Nucleic Acids Res. *39*, 1647–1655.
87. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). 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*, 405–423.
88. Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res. *34*, D535–D539.
89. Barber, L.J., Youds, J.L., Ward, J.D., McIlwraith, M.J., O'Neil, N.J., Petalcorin, M.I.R., Martin, J.S., Collis, S.J., Cantor, S.B., Auclair, M., et al. (2008). RTEL1 Maintains Genomic Stability by Suppressing Homologous Recombination. Cell *135*, 261–271.
90. Uringa, E.-J., Lisaingo, K., Pickett, H. a, Brind'Amour, J., Rohde, J.-H., Zelensky, A., Essers, J., and Lansdorp, P.M. (2012). RTEL1 contributes to DNA replication and repair and telomere maintenance. Mol. Biol. Cell *23*, 2782–2792.
91. Berman, H.M. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.
92. Pieper, U., Webb, B., and Dong, G. (2014). ModBase, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res.
93. Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2014). Ensembl 2015. Nucleic Acids Res. *43*, D662-669.

94. The UniProt Consortium (2014). UniProt: a hub for protein information. Nucleic Acids Res. *43*, D204-212.
95. Velankar, S., Dana, J.M., Jacobsen, J., van Ginkel, G., Gane, P.J., Luo, J., Oldfield, T.J., O'Donovan, C., Martin, M.-J., and Kleywegt, G.J. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res. *41*, D483-9.
96. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.
97. Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. *48*, 443–453.
98. Gaines, K.F., Bryan, a L., and Dixon, P.M. (2000). The Effects of Drought on Foraging Habitat Selection of Breeding Wood Storks in Coastal Georgia. Waterbirds *23*, 64–73.
99. Diggle, P.J., and Chetwynd, a G. (1991). Second-order analysis of spatial clustering for inhomogeneous populations. Biometrics *47*, 1155–1163.
100. Pieper, U., Eswar, N., Davis, F.P., Braberg, H., Madhusudhan, M.S., Rossi, A., Marti-Renom, M., Karchin, R., Webb, B.M., Eramian, D., et al. (2006). MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res. *34*, D291-5.
101. Turner, T.N., Douville, C., Kim, D., Stenson, P.D., Cooper, D.N., Chakravarti, A., and Karchin, R. (2015). Proteins linked to autosomal dominant and autosomal recessive disorders harbor characteristic rare missense mutation distribution patterns. Hum. Mol. Genet. *24*, 5995–6002.
102. Stenson, P.D., Ball, E. V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S.T., Abeysinghe, S., Krawczak, M., and Cooper, D.N. (2003). Human Gene Mutation Database (HGMD): 2003 Update. Hum. Mutat. *21*, 577–581.
103. Futreal, P., Coin, L., Marshall, L., Down, T., Hubbard, T., Wooster, T., Rahman, N., and Stratton, M. (2004). A census of human cancer genes. Nat Rev Cancer *4*, 177–183.
104. Chakravarty, D., Gao, J., Phillips, S., and Kundra, R. (2017). OncoKB: a precision oncology knowledge base. Precis. Oncol. *[epub]*,.
105. Tartaglia, M., Mehler, E.L., Goldberg, R., Zampino, G., Brunner, H.G., Kremer, H., van der Burgt, I., Crosby, a H., Ion, A., Jeffery, S., et al. (2001). Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. Nat. Genet. *29*, 465–468.
106. Kontaridis, M.I., Swanson, K.D., David, F.S., Barford, D., and Neel, B.G. (2006). PTPN11 (Shp2) mutations in LEOPARD syndrome have dominant negative, not activating, effects. J. Biol. Chem. *281*, 6785–6792.
107. Lahiry, P., Torkamani, A., Schork, N.J., and Hegele, R. a (2010). Kinase mutations in human disease: interpreting genotype-phenotype relationships. Nat. Rev. Genet. *11*, 60–74.
108. Hall, J.B., Dumitrescu, L., Dilks, H.H., Crawford, D.C., and Bush, W.S. (2014). Accuracy of administratively-assigned ancestry for diverse populations in an electronic medical record-linked biobank. PLoS One *9*, 1–6.
109. Lee, S., Fuchsberger, C., Kim, S., and Scott, L. (2016). An efficient resampling method for calibrating single and gene-based rare variant association analysis in case-control studies. Biostatistics *17*, 1–15.

110. Blake, J.A., Eppig, J.T., Kadin, J.A., Richardson, J.E., Smith, C.L., Bult, C.J., Anagnostopoulos, A., Baldarelli, R.M., Beal, J.S., Bello, S.M., et al. (2017). Mouse Genome Database (MGD)-2017: Community knowledge resource for the laboratory mouse. Nucleic Acids Res. *45*, D723–D729.
111. Dewey, F.E., Murray, M.F., Overton, J.D., Habegger, L., Leader, J.B., Fetterolf, S.N., O'Dushlaine, C., Van Hout, C. V., Staples, J., Gonzaga-Jauregui, C., et al. (2016). Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science (80-.). *354*, aaf6814.
112. Samocha, K.E., Kosmicki, J.A., Karczewski, K.J., O'Donnell-Luria, A.H., Pierce-Hoffman, E., MacArthur, D.G., Neale, B.M., and Daly, M.J. (2017). Regional missense constraint improves variant deleteriousness prediction. bioRxiv.