FUNCTIONAL ASSESSMENT OF AMINO ACID VARIATION IN HUMAN GENOMES

A Dissertation
Presented to
The Academic Faculty

By

Thanawadee Preeprem

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics

Georgia Institute of Technology

May 2014

FUNCTIONAL ASSESSMENT OF AMINO ACID VARIATION IN HUMAN GENOMES

Approved by:

Dr. Greg Gibson, Advisor
School of Biology
Georgia Institute of Technology

Dr. Stephen C. Harvey
School of Biology
Georgia Institute of Technology

Dr. I. King Jordan
School of Biology
Georgia Institute of Technology

Dr. Fredrik O. Vannberg
School of Biology
Georgia Institute of Technology

Dr. May D. Wang
Dept. of Biomedical Engineering
Georgia Institute of Technology

Date Approved: March 27, 2014

To my parents

ACKNOWLEDGEMENTS

I have been fortunate and grateful to have Dr. Greg Gibson as my advisor. I joined his lab late in my graduate life and he provided me an excellent research environment that I can discuss my ideas with him enthusiastically. I appreciate very much of his valuable scientific advice and his great help that has made my dissertation possible.

I also would like to thank Dr. Stephen C. Harvey, my former advisor, for his help in my personal and academic development. I am grateful for his teachings and trainings that have made me an independent and very capable researcher.

The mentorship from my both advisors has contributed a lot to my success.

I appreciate the guidance, kind advice and feedback from my committee members: Dr. Greg Gibson, Dr. Stephen C. Harvey, Dr. I. King Jordan, Dr. Fredrik O. Vannberg, and Dr. May D. Wang. Their encouragements also boost up the confident in what I do. All of their advices will continue to motivate me during my future academic career.

During my teaching assistantship, I had privileges to work with three distinct professors: Dr. Mark Borodovsky, Dr. Ingeborg Schmidt-Krey, and Dr. Mirjana Brockett. Working with them helped me improve a great deal of understanding of academic life. I enjoyed teaching and was delighted to receive great feedbacks from these role models. The experience I gained had sharpened my teaching skills till these days.

I want to thank all of the past and present members of the Harvey's and Gibson's labs for their friendship and collaborations. My special thanks go to Minmin Pan, Burak Boz and,

Jared Gossett from the Harvey's lab for everything they shared with me in my early Ph.D. career.

In addition to those individuals, I also have come across many more people during my graduate study. I acknowledge their lively and inspiring scientific conversations which keep my fascination for science alive.

I am very pleased to have Ziming Zhao, Jianrong Wang, Burak Boz, Jared Gossett, and Shefaet Rahman as friends who are always with me in every aspect of my ups and downs.

I appreciate my parents, sister and brother for their love, support, help, and patient. For me, pursuing a Ph.D. is quite a difficult and a lonely path. My family's belief in me always keeps me focused and be positive.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

ΔΔG	Gibbs free energy change
3D	3-dimensional
Å	Ångstrom
AA	African American
AACDS	Association-Adjusted Consensus Deleterious Scheme
B-factor	Temperature factor
CHDWB	Center for Health Discovery and Well Being
Con count	Conservation count
C_α	Central carbon atom in an amino acid
dbNSFP	Database for nonsynonymous SNPs' functional predictions
dbSNP	Single nucleotide polymorphism database
Del count	Deleterious count
EA	European American
GWAS	Genome-wide association study
MAF	Minor allele frequency
MSV3d	Database of human MisSense Variants mapped to 3D protein structure

nsSNP	Nonsynonymous Single Nucleotide Polymorphism
PDB	Protein Data Bank
Phyre2	Protein Homology/AnalogY Recognition Engine 2
RMSD	root-mean-square deviation
RSA	Relative solvent accessibility
SAHG	Structural Atlas of Human Genome
SC	Stabilization center
SDS	Structural Disruption Score
SNP	Single Nucleotide Polymorphism
SR	Stabilizing residue
SwissVar	Portal to Swiss-Prot diseases and variants
UniProt	Universal Protein Resource
VIP	Very Important Pharmacogene

SUMMARY

The Human Genome Project, initiated in 1990, creates an enormous amount of excitement in human genetics—a field of study that seeks answers to the understanding of human evolution, diseases and development, gene therapy, and preventive medicine. The first completion of a human genome in 2003 and the breakthroughs of sequencing technologies in the past few years deliver the promised benefits of genome studies, especially in the roles of genomic variability and human health. However, intensive resource requirements and the associated costs make it infeasible to experimentally verify the effect of every genetic variation. At this stage of genome studies, *in silico* predictions play an important role in identifying putative functional variants.

The most common practice for genome variant evaluation, the sequence conservation-based approach, assumes important positions in a DNA or a protein sequence have been conserved throughout the evolution. Therefore, the predicted variant effect (deleterious or benign) is based on the evolutionary conservation at the mutation site. Nonetheless, phylogenetic diversity of aligned sequences used to construct the prediction algorithm has substantial effects on the analysis since sequence conservation is not the absolute predictor for deleteriousness.

This dissertation aims at overcoming the weaknesses of the conservation-based assumption for predicting the variant effects. The dissertation describes three different integrative computational approaches to identify a subset of high-priority amino acid mutations, derived from human genome data.

For genetic variants found in genomes of healthy individuals, an eight-level Association-Adjusted Consensus Deleterious Scheme (AACDS) is implemented. It ranks amino acid mutations based on suggestive evidence from association studies and conservation-based predictors. The ranking scheme has particular utility for the development of individualized health profiles. A database-driven web application promotes the utility of AACDS by granting access to AACDS data for 68 million amino acid mutations in over 18,000 human genes.

For candidate genetic variants of epilepsy disorders, a novel 3-dimensional structure-based assessment protocol for amino acid mutations is established. It models protein structure and contrasts structural characteristics that distinguish between residues altered by disease-causing mutations vs. neutral mutations. A structural disruption score (SDS) is introduced as a measure to depict the likelihood that a candidate variant is functional. SDS is correlated with standard conservation-based deleteriousness, but shows promise for improving discrimination between neutral and causal variants at less conserved sites. The implementation facilitates variant prioritization for experimental validations.

For genomic variants that may affect inter-individual variability in drug responses, a systematic 3-dimensional screening is performed on key proteins of drug metabolism. “SDS Pharmacogenes” is an explicit structure-based predictor that comprises five predictive features for structural disturbances. Unlike conservation-based predictors, SDS Pharmacogenes is able to annotate VIP variants as functional rather than neutral mutations based on the distinguishable characteristic profiles of structural disturbances. The systematic variant evaluation pipeline allows efficient structural examination of multiple variants, thus provide an opportunity to investigate their joint effects. The

implementation of SDS Pharmacogenes and the future development of a database-driven web application will offer comprehensive understanding of genetic variants and introduces a new approach for aiding optimization of drug therapy

CHAPTER 1: INTRODUCTION

Genomic variations of human genomes

Genetic variation refers to differences in DNA of a person when compared to the DNA of a human reference genome. Genome studies indicate that each healthy individual carries a large number of genetic variants (**Figure 1.1**) [1-4]. Single nucleotide variants (SNVs) are the most common form. Single nucleotide polymorphisms (SNPs), a subset of SNVs, are single base genetic variations observed in at least 1% of the population. Both SNVs and SNPs can be associated with diseases. SNVs may alter gene expression levels via regulatory controls. Nonsynonymous SNPs (nsSNPs), also known as missense variants, are SNPs that introduce amino acid changes to proteins. The change may affect protein function. Both types of genetic variation contribute significantly to disease states: about 60% and 25% of known disease-causing mutations are nsSNPs and regulatory SNVs, respectively [5].

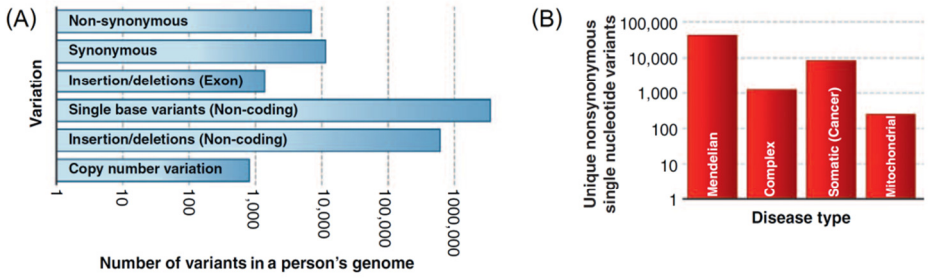


Figure 1.1: Types and number of genetic variations found per a human genome and numbers of known nsSNPs that are associated with diseases. (adapted from [1])

Current challenges in genome studies

Genome studies have several goals, one of which is to identify genes and genetic variants that give rise to diseases. Scientists are now facing several challenges in genome studies [6, 7], particularly in the area of identifying candidate genes and/or causal mutations. To infer candidate genes, the study requires complete re-sequencing of candidate genomic intervals in large population samples [6]. To test the hypothesis that a given SNP is a causative factor, the process requires: (1) linkage-disequilibrium studies on a very dense SNP map, and (2) comparisons of allele frequencies between affected and unaffected individuals [8]. Nevertheless, genetic associations between case and control groups may also be due to several discrepancies, most of which are functionally irrelevant. In addition, the direct method requires follow-up studies to further examine the genetic differences: typically by surveying larger panels of control individuals for the presence or absence of exclusive case SNPs [9].

The confirmation of true positives for disease-causing mutations entails laboratory testing [10]. The process is difficult and expensive. When a genetic variant is found in a well-known disease-causing gene, causality is difficult to conclude. For genes without known functions, functional assays are not easy to design [6]. Taking into account that some researchers do not publish results of failed attempts, a rough estimate on the number of truly causative variants from exome sequencing projects is 10-50% of cases [6]. Despite the difficulties in establishing causality, the scientific community still shows great interest in pursuing genome studies to understand the core of human health and the possible discovery of cures for diseases.

Roles of bioinformatics for evaluating genome variants

Major efforts have been made in recent years to identify nsSNPs (missense variants) with strong effects because of their prevalence in number and high penetrance of association with diseases. Intensive resource requirements and the associated cost make it unfeasible to verify the effect of every genetic variation. At this stage of human genome study, *in silico* predictions play an important role in identifying putative functional variants in a systematic and efficient manner. Computational tools substitute the current elaborate process of *in vitro* and *in vivo* experiments. Several approaches have been developed to narrow down the list of candidate variants or genes from genome data [11-13]. Irrespective of differences in underlying algorithms and scoring functions, most tools evaluate nsSNPs as either deleterious (having a strong functional effect) or neutral (having a weak functional effect) from the level of DNA or protein sequence conservation [14]. The results often indicate a probability that the amino acid substitution is damaging.

Recently, a database named dbNSFP [15] has compiled prediction scores of commonly used deleterious predictors into a stand-alone database for easy retrieval of scores. This contribution greatly facilitates the evaluation of variant deleteriousness in large genome datasets. Scores from six prediction algorithms, with various complimentary methodologies were included. They are: conservation-based (SIFT [16], LRT [17], and MutationAssessor [18]), conservation- and structure-based (PolyPhen2_HumDiv and PolyPhen2_HumVar [19]), and the Bayesian classifier MutationTaster [20].

Limitations of sequence conservation-based approaches for predicting variant deleteriousness

While a conservation-based approach is the common practice for assessing the effects of nsSNPs, there are still some concerns when it comes to interpreting the results: (1) Sequence conservation is not an absolute predictor for deleteriousness— phylogenetic diversity of sequences used to infer evolutionary conservation has substantial effects on the analysis [14]. (2) The score cut-offs for different levels of variant deleteriousness are sometimes not well determined [21]. (3) Predictions from several programs occasionally differ; it is quite problematic to represent a consensus result [22]. (4) Some documented disease-causing variants may be false positives—prediction programs trained on this data set may suffer from the classification error [6]. (5) Some novel variants may be misclassified since there is not enough phenotypic information to infer the significant effects of nsSNPs [23]. (6) A predicted deleterious variant may result in reduced or damaged protein functions; however, the chance that the altered phenotype is observable may be small. (7) Variant classification programs which are mature enough for general use may have low accuracy for certain genes [24, 25].

Integrative analysis of variant effects: a novel approach for personal genome interpretations

Because of advances in sequencing technologies [26], personal genome analysis is now being marketed for evaluating disease risk and/or adverse drug responses in healthy individuals, or for identifying disease-causing variants in people with diseases. Genome interpretation is a complex and time-consuming process. This requires a good knowledge of genetic principles and in-depth evaluation of genetic variation. Even if existing

deleterious prediction algorithms show a reasonable range of accuracy (~70-80%) [22, 27] for evaluating nsSNPs, the lack of comprehensive knowledge of protein biology and its role in human disease restricts the usage of prediction results in the practical way.

Information from three sources (clinical annotations, protein sequences, and protein structures) can be used simultaneously to estimate the functional significance of a variant [28]. The integrative approach utilizes prior knowledge of genome biology to define a set of genetic variants that are important for an individual's health. Researchers have begun to apply this integrative procedure in variant deleteriousness algorithms [29-32], but the implementations have not yet realized their full potential. The reasons include: the massive amount of genomic-related data yet to be explored, the insufficient expertise to perform certain analyses, and sub-optimal procedures for combining prediction results from various predictors. Once the integrative protocol is mature, it will have great impact on personal genome interpretation and variant prioritization; that is, promoting the comprehensive understanding of genomic variants from interrelated SNP data.

1) Adding clinical annotations to conservation-based approaches

Owing to the rapid improvement of sequencing technologies, data on the clinical associations with a SNP, a protein residue, or a gene are accessible in many SNP databases [11, 13, 30, 33]. However, these functional annotations are quite scattered; the search for the clinical effects of a variant may be time-consuming and as a result, obstruct the practical analysis of genomic variations. The meaningful combination of interrelated data and the extension of existing predictive features to include any prior knowledge will support the ease and comprehensive evaluation of genomic data. This remark is

addressed in the implementation of the Association-Adjusted Consensus Deleterious Scheme (AACDS) for categorizing personal genome variations (**Chapter 2**).

2) Incorporating protein structures into conservation-based approaches

An amino acid mutation is caused by a single point mutation of the coding nucleotide. Many amino acid pairs have different physicochemical properties, so the mutant residues may prompt observable effects on protein functions. The consequence of amino acid mutations can be readily predicted by many algorithms, based on protein sequences or structures [12, 34]. Some genomic tools have begun to incorporate protein analysis algorithms into the traditional conservation-based predictors [13, 35-39].

Visual inspections of wild type vs. mutant protein structures can help pinpoint the protein residues that are responsible for aberrant protein functions. Success in homology modeling alleviates the fundamental limitation of structure-based approach; i.e., the limited availability of 3-dimensional (3D) protein structures. Using various sources for protein structures, my Structural Disruption Score (SDS) is established. It represents predictive features for amino acid mutations that induced by the candidate case missense variant for epilepsy disorders (**Chapter 4**) and by missense variants in pharmacogenes (**Chapter 5**).

Description of this dissertation

My dissertation aims to overcome the weaknesses of traditional conservation-based predictions for variant deleteriousness. The integration of conservation- and structural-based approaches with database searches and literature surveys provides a better

understanding of protein function, enables the identification of functional protein residues, and promotes a more accurate classification of genetic variants with respect to overall protein function or disease development. The implementation is applied to three disciplines of genome analysis, and restricted to the evaluations of amino acid mutations induced by missense variants (nsSNPs) since they are the most prevalent type of genetic variations that are closely associated with diseases [5].

The first objective is to categorize personal genome variations of healthy individuals into distinct classes; each class of nsSNPs reflects the strength of evidence which collectively suggest the variants may contribute to adverse gene functions. The implementation of this novel variant prioritization scheme, an Association-Adjusted Consensus Deleterious Scheme (AACDS), is described in **Chapter 2**. To promote the utility of this schema, a database-driven web application is developed (**Chapter 3**).

The second objective is to develop a disease-specific predictor to evaluate missense variants found in a set of epilepsy-related genes (**Chapter 4**). This implementation integrates results of conservation-based predictors with the outputs of many protein structural analyses. Distinguishable characteristics between documented disease-causing and neutral mutations within these genes are used to establish a "Structural Disruption Score" (SDS) for epilepsy variants. Using SDS, the candidate variants are ranked based on the likelihood that each mutation may disrupt protein structure. The list of high priority variants can be used to assist further experimental validations.

The last objective is to extend the implementation of SDS into an exclusive structure-based predictor for evaluate missense variants in pharmacogenes (**Chapter 5**). The "SDS

Pharmacogenes" consists of 13 predictive features; it can be used to evaluate almost every mutated residue within the 45 gene products.

CHAPTER 2: AN ASSOCIATION-ADJUSTED CONSENSUS DELETERIOUS SCHEME TO CLASSIFY HOMOZYGOUS MISSENSE MUTATIONS FOR PERSONAL GENOME INTERPRETATION [40]

Abstract

Background: Personal genome analysis is now being considered for evaluation of disease risk in healthy individuals, utilizing both rare and common variants. Multiple scores have been developed to predict the deleteriousness of amino acid substitutions, using information on the allele frequencies, level of evolutionary conservation, and averaged structural evidence. However, agreement among these scores is limited and they likely over-estimate the fraction of the genome that is deleterious.

Method: This study proposes an integrative approach to identify a subset of homozygous non-synonymous single nucleotide polymorphisms (nsSNPs). An 8-level classification scheme is constructed from the presence/absence of deleterious predictions combined with evidence of association with disease or complex traits. Detailed literature searches and structural validations are then performed for a subset of homozygous 826 missense mutations in 575 proteins found in the genomes of 12 healthy adults.

Results: Implementation of the Association-Adjusted Consensus Deleterious Scheme (AACDS) classifies 11% of all predicted highly deleterious homozygous variants as most likely to influence disease risk. The number of such variants per genome ranges from 0 to

8 with no significant difference between African and Caucasian Americans. Detailed analysis of mutations affecting the APOE, MTMR2, THSB1, CHIA, αMyHC, and AMY2A proteins shows how the protein structure is likely to be disrupted, even though the associated phenotypes have not been documented in the corresponding individuals.

Conclusions: The classification system for homozygous nsSNPs provides an opportunity to systematically rank nsSNPs based on suggestive evidence from annotations and sequence-based predictions. The ranking scheme, in-depth literature searches, and structural validations of highly prioritized missense mutations compliment traditional sequence-based approaches and should have particular utility for the development of individualized health profiles. An online tool reporting the AACDS score for any variant is provided at the authors' website.

Background

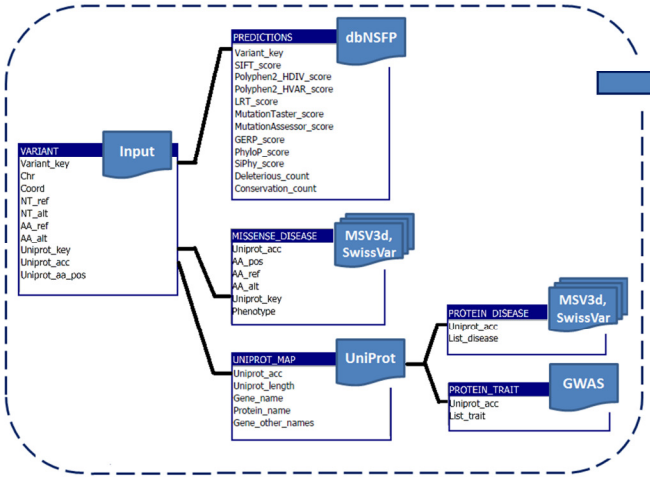
Personal genome interpretation is a process of determining the personal genome sequences and assessing the likely consequences of an individual's genetic variation. Personalized genome data interpretation can be used, for example, to predict diseases and traits, identify mutations for family planning purposes, and guide medical treatments based on likely drug responses. Developments in next-generation sequencing technologies over the past five years have enabled personal genome interpretation to become feasible and affordable [26]. Despite these advances, understanding of the impact of specific genetic variants remains limited. Major efforts have been made to identify nsSNPs with strong effects because of their collective high prevalence and likelihood that many may be clinically actionable [5].

Sequence-based prediction algorithms are commonly used to categorize nsSNPs into damaging and non-damaging, and to predict the effects (small or large) of nsSNPs with respect to undesirable phenotypes. The algorithms score the amino acid changes from the level of sequence conservation observed in homologous sequences or from the degree of physicochemical changes. Structure-based predictions evaluate 3-dimensional (3D) structural features, e.g., solvent accessibility, stability, number of residue contacts, which are altered in the mutant proteins. Approaches that include annotation of biological function also support functional assessments of amino acid substitutions. Ng and Henikoff (2006) proposed that the combination of all three types of data may provide the most accurate assessment of likely deleteriousness [28], which motivates the development of the schema proposed in this study.

A pioneer study in personal genome interpretation stated that the human reference genome carries 1104 nsSNPs *predicted* to have impact on protein functions [41]. A similar study indicated that there are 796–837 *predicted* deleterious nsSNPs per individual [17]. This number of *predicted* damaging nsSNPs is much greater than both the *theoretically estimated* 15–60 damaging nsSNPs per genome [5] and the *classified* disease-causing nsSNP number of 40–100 per genome [42]. These observations highlight the complexity of personal genome interpretation and the need for a variant classification schema that builds on algorithmic prediction by integrating sound knowledge of the biological and structural impact of genetic variants.

There are many databases that provide useful information about genetic variants. Because genetic polymorphisms found in healthy individuals tend to have small effects, further

improvement of available resources is required to more accurately define the set of variants that are likely to be most important for an individual's health. In this study, we constructed a classification schema (**Figure 2.1**, **Supplementary Figure A.1**) to rank nsSNPs identified in healthy individuals by their functional significance. Each ranking category reflects the strength of evidence that a variant may adversely affect gene function from several standpoints, incorporating both database searches and sequence-based predictions. The newly developed variant classification scheme is designed to generate a best estimate of clinical significance for each variant of interest, with the intention of focusing attention on the most likely deleterious SNPs.



Category	Features of SNP			
	Disease-causing	Deleterious count ≥ 3	In disease genes	In trait genes
1	X			
2A			X	
2B		X	X	
3A				X
3B		X		X
4		X		
5			X	X
6				

Figure 2.1: Flow diagram for AACDS classification algorithm. Upon receiving a list of homozygous rare missense variants, the nsSNPs were mapped to the corresponding amino acid residue within a reference protein (identified with UniProt accession number). We use an SQL server to hold SNP-related information from several resources: deleterious predictions (from dbSNFP database [15]), known diseases associations of each variant and known disease/trait associations of each gene (from MSV3d [43], SwissVar [44], and GWAS databases [45]). The AACDS classification algorithm extracts relevant fields from the AACDS database to populate the report on the variant category (categories 1–6). The output (**Supplementary Figure A.1**) is converted from an SQL data table to the user-defined formats (HTML, text), and is available for download, or individual queries can be supported on our server at http://www.cig.gatech.edu/tools.

Given the very large number of candidate disease-promoting variants per genome, we here focus just on the homozygous variants reasoning that highly penetrant effects are most likely to be recessive. The methods developed could be applied to all heterozygous nsSNPs as well, but this would be a daunting task for manual inspection, which would only be warranted given extensive phenotype data and a desire of an individual to receive the information. Here we describe homozygous nsSNPs in the genomes of 12 healthy participants in a predictive health study, the Emory-Georgia Tech Center for Health Discovery and Well Being (CHDWB). Since the IRB consent does not allow communication of genetic data, given potential negative consequences of knowledge of variants that cannot be acted upon, the identities of the individuals are anonymous and no concerted attempt has been made here to link genotypes to phenotypes directly. Expanded and appropriately consented studies will be required to evaluate the actual utility of the proposed schema as a means of focusing attention on those variants that are most likely to influence personalized health behaviors.

Methods

Whole genome sequence dataset

Whole genome sequence (WGS) data was obtained for 12 healthy adult participants in the Center for Health Discovery and Well Being (CHDWB) study, including 4 African American women, 4 Caucasian women, and 4 Caucasian men. None of the individuals has any known complex or Mendelian diseases, but they cover a variety of profiles with respect to overall physical and mental health. Prediction of disease risk based on common variants and clinical profiles is described for the Caucasians in [46]. The participants

have provided written consent to publication of their whole genome sequence data for research purposes only. They do discuss their clinical profile with a health professional following annual visits to the Center, but are not currently permitted to receive the genetic data generated during the study. In order to protect participant identities, their identifying numbers have been randomized for this study.

WGS was performed on Illumina HiSeq2000 automated sequencers at the University of Washington facility under contract to the Illumina Genome Network. Approximately 125 billion bases passing the Illumina analysis filter were obtained for each genome. Mean non-N reference coverage (after excluding gaps) is ~36X with 95.5% of the positions having an average coverage of at least 10X. The genome sequences were aligned against the human reference genome assembly (hg19 sequence) using CASAVA (Consensus Assessment of Sequence And VAriation, Illumina, Inc, San Diego, CA). On average, 87% of each individual's quality filtered reads were aligned. High-confidence variants with a quality score above 20 were retained. Accuracy of the generated genome sequences was confirmed by comparison with previously determined genotypes from Illumina OmniQuad arrays, which showed >99% concordance for all individuals.

The coding variants were functionally annotated using Variant Annotation Tool (VAT) [47] which uses the GENCODE v7 gene annotation set [48]. We identified "homozygous rare variants" with allele frequency <10% in the Caucasian or African 1000 Genomes dataset, using the data provided by dbNSFP [15]. All homozygous rare nsSNPs relative to hg19 were identified for each person, and assigned to two categories: known nsSNPs that are present in dbSNP build 137 [49], and private nsSNPs that are absent from dbSNP

but found exclusively in each individual. Information on the number of variants of each type is provided in Table 1. Minor allele frequencies (MAFs) for all nsSNPs were obtained from NHLBI GO Exome Sequencing Project (ESP6500) (June 2012 release) [50], Genomic data for known homozygous nsSNPs (n = 826, including 29 private variants) were analyzed as a whole and per individual.

Amino acid indices for the alternate residues were mapped to the corresponding proteins using transcript IDs for the major isoform. All protein sequences and related information including protein functions and sequence features were obtained from the UniProt database [51].

Table 2.1: Summary of genetic variations in genome sequences of 12 individuals.

Subject ID	Eth, sex	Total Variants (>q20)			Coding variants (based on Gencode v7)								Homozygous nsSNPs (based on dbSNP build 137)			
					SNPs				Indels			SVs				
		SNPs	Indels	SVs	S	MS	NS	Splice Overlap	Indels FS	Indels NFS	Indels Overlap	SVs Overlap	#Known nsSNPs	#Unique genes	#*de novo* nsSNPs	#Unique genes
1	Afr, F	4513763	733596	4251	14793	14039	72	98	381	342	137	37	88	77	2	2
2	Afr, F	4472988	754399	4545	14500	13712	66	106	393	335	147	55	71	63	3	3
3	Afr, F	4287739	722922	4447	13755	13166	84	79	374	301	120	43	77	71	3	2
4	Afr, F	4443799	746111	4368	14488	13874	73	104	366	338	142	40	58	56	1	1
5	Cau, F	3734820	645032	3977	11929	11745	62	90	343	307	123	43	57	40	2	2
6	Cau, F	3691337	633475	4114	11757	11457	56	90	317	280	106	49	52	45	none	none
7	Cau, F	3691270	632544	4033	11912	11488	65	71	279	304	116	37	50	44	4	4
8	Cau, F	3722234	641792	4197	11887	11434	64	76	303	299	125	41	55	42	none	none
9	Cau, M	3647944	590064	3828	11619	11255	54	76	311	281	95	38	60	50	2	2
10	Cau, M	3643046	597363	4011	11814	11480	61	85	289	287	109	31	82	65	2	2
11	Cau, M	3650690	602744	3916	11560	11285	60	80	342	280	112	32	72	65	5	5
12	Cau, M	3701558	639005	4739	11842	11708	60	81	334	290	118	37	75	64	5	5
													#Total = 797	**#Unique = 575**	**#Total = 29**	**#Unique = 25**

Abbreviations: *Eth* ethnicity, *SNPs* single nucleotide polymorphisms, *SV* structural variants, *S* synonymous, *MS* mis-sense, *NS* nonsense, *FS* frameshift, *NFS* non frameshift.

Sequence annotation using published algorithms

Evidence for association of each SNP or gene with diseases or traits was obtained from public repositories of amino acid polymorphisms (MSV3d, July 2012 release [43] and SwissVar, accessed December 2012) [44], from Online Mendelian Inheritance in Man (OMIM) [52], and from the NHGRI genome wide association studies (GWAS) [45] catalog. Initially, each nsSNP was assigned as disease-causing, probably disease causing, unclassified, or neutral. Functional predictions, and information on disease- and trait-associations to the gene were collected from dbNSFP [15]. In addition, we used UniProt sequence feature records [51] to annotate whether the mutated amino acid is localized to any structurally/functionally important sites (molecule processing sites, binding sites, modification sites, etc.).

To annotate deleterious nsSNPs, consensus predictions from several algorithms were compared. Pre-computed deleterious scores for each nsSNP were retrieved from dbNSFP [15]. To our knowledge, dbNSFP is the first database that provides pre-computed functional predictions from multiple algorithms, facilitating interpretation of the deleteriousness of variants in large datasets. The database provides the output of six different prediction algorithms that have complimentary methodologies. Three are sequence-based (SIFT [16], LRT [17], and MutationAssessor [18]), while two are both sequence and structure-based (PolyPhen2_HumDiv and PolyPhen2_HumVar [19]), and the sixth is the MutationTaster Bayesian classifier [20]). Each of these tools relies primarily on the basic assumption that residue functionality dictates sequence conservation, which can consequently be used to infer deleteriousness. Raw scores for the first five programs were re-scaled to [0, 1] in which a score closer to 1 represents a

stronger (deleterious) effect of a nsSNP [15]. A MutationAssessor score of > 3.5 designates high functional impact [18], hence, “deleterious”. The six prediction programs were used to construct the classification scheme of nsSNPs presented in **Table 2.2**. Putative deleterious nsSNPs were identified as nsSNPs reported as “deleterious” by at least three out of six prediction programs.

Table 2.2: AACDS classification of homozygous nsSNPs in 12 genomes.

Category	Features of SNP				Subject ID, Ethnicity, Sex												Total #nsSNPs
	Disease-causing	Deleterious count ≥ 3	In disease genes	In trait genes	1	2	3	4	5	6	7	8	9	10	11	12	
					Afr, F	Afr, F	Afr, F	Afr, F	Cau, F	Cau, F	Cau, F	Cau, F	Cau, M	Cau, M	Cau, M	Cau, M	
1	X				1	none	none	none	1	none	none	1	1	none	1	none	**5**
2A			X		8	7	6	5	6	5	11	7	12	7	10	9	**93**
2B		X	X		1	none	none	none	2	2	4	1	2	none	2	none	**14**
3A				X	9	24	17	7	6	10	9	10	5	13	16	10	**136**
3B		X		X	1	5	none	none	1	1	3	4	2	2	none	2	**21**
4		X			5	5	8	6	none	5	3	7	4	6	5	4	**58**
5			X	X	14	21	20	10	8	12	9	11	13	18	20	14	**170**
6					68	40	49	42	46	32	33	33	40	56	45	55	**539**
Total #unique nsSNPs					**88**	**71**	**77**	**58**	**57**	**52**	**50**	**55**	**60**	**82**	**72**	**75**	**797**

The number of homozygous nsSNPs per individual and the number of SNPs in each AACDS category are specified.

In addition to the deleterious predictions, protein regions under evolutionary constraint were detected using three evolutionary conservation-based indicators: GERP++ [53], phyloP [54], and SiPhy [55]. We also used Grantham scores [56] to reflect the degree of physicochemical differences between pairs of amino acids. These four indicators were included in the analysis for comparison purposes but were not utilized for nsSNP categorization.

Other popular variant annotation tools might also be useful but were considered to be redundant with respect to our purposes. For example, ANNOVAR [57] has the ability to perform variant annotation (intronic, intergenic, untranslated region, exonic: non-synonymous, synonymous, etc.), but this information was already available from the VAT output. At the time of our analysis, ANNOVAR provided dbSNP build 135 mapping, not the dbSNP build 137 [49] used in our analysis pipeline. Note that gene definitions from ANNOVAR refer to the nucleotide reference sequence where a SNP is located; the format is not directly applicable for working with records from the two selected SNP databases (MSV3d [43] and SwissVar [44]), in which UniProt accessions were used to identify gene products. Similar to ANNOVAR, SnpEff [58] is another popular program that can assign structural annotations of variants. This function would be valuable when one wants to analyze different types of genetic variations within a genome. Because we only focused on analysis of missense mutations, the annotation feature of SnpEff was deemed unnecessary.

Supervised and automated structure-based predictions of variant function

High quality protein 3D structures are essential to identify functional impacts of nsSNPs. Due to the limited availability of experimentally-determined human protein structures [59], an assortment of 3D structure sources was used to manually evaluate the effects of single point mutations found in specific proteins (**Supplementary Table A.1**). Crystal structures were retrieved from the RCSB Protein Data Bank (PDB) [60]. Homology models were retrieved from Protein Model Portal (PMP) [61] repository, or were built manually by joining multiple structures into a single model of a protein. Steric conflicts found within homology models were resolved by energy minimization with explicit solvent using YASARA force field [62]. Structural validation of homology models was evaluated by using two independent scores: QMEAN6 [63] and ModFOLD4 [64]. All 3D structures were visualized and rendered using Chimera [65].

The analysis began with a visualization of wild type proteins in the context of bound ligands. Additional variants that are known to be associated with diseases, or affect protein functionality and/or stability were also identified for each protein structure. Next, we used SDM [66] to compare the protein stability changes upon amino acid mutations with the default modeling of a mutant structure using Andante [67]. A mutation is classified as affecting protein function (stabilizing or destabilizing) using the stability cut-off of ±2 kcal mol^{-1} [66]. For evaluating the impact of amino acid changes on protein stability in a high throughput fashion, we obtained the tertiary classification of protein stability changes (increase, decrease, neutral) caused by a SNP from I-Mutant 2.0 [68], available from MSV3d [43]. The predictions are based on the protein (or homolog) structure or solely on the protein sequence when the structure was unavailable.

In order to expand the structure-based predictions to a larger dataset without the requirements of manual inspection or in-depth literature searches, we applied a combination of database searches and computational predictions to a larger set of proteins. In addition to the 6 protein structures described in the main text, we identified an additional 25 protein coordinates from PDB [60] (**Supplementary Table A.2**). We assessed these 25 wild type proteins in 4 areas of structural analysis: protein stability, ligand binding capability, protein dynamics, and protein-protein interactions. For protein stability, we used the aforementioned approach along with predictions of amino acids with specialized roles regarding protein stability, namely long-range stabilization center (SC) residues and stabilizing residues (SRs). These residues were inferred from the SCide [69, 70] and SRide webservers [71], respectively. Ligand binding residues for each protein were retrieved from PDBe (www.ebi.ac.uk/pdbe/) or were predicted using 3DLigandSite [72]. Amino acid residues that are located in or near predicted binding pockets are likely to alter the binding capability for ligand(s). As disease-causing mutations that do not occur in binding sites or buried sites are predominantly found on protein interfaces [36], we used the PatchFinder program [73, 74] to computationally predict the most significant cluster of conserved residues on a protein's surface that may indicate possible functional sites of the protein; i.e., sites of protein-protein interactions. Changes in protein dynamics were evaluated by the crystallographic B-factor of Cα atoms. In addition, we also used PredyFlexy [75] and FlexPred [76, 77] to predict the dynamic class of an amino acid residue (rigid, intermediate, flexible), and to estimate whether each residue is likely to induce conformational switches within the protein.

Assessment of functional enrichment

The g:Profiler web server [78] was used to detect enrichment of gene functions for genes whose nsSNP are homozygous. Functional profiling and statistical enrichment analysis were performed with two distinct methods. First, we compared the annotations of multiple gene lists, where each list represents the genes with known homozygous variants found in an individual, using G:Cocoa (n = 40–77 genes per genome). Then, we analyzed a gene list for each individual using G:GOSt. The second analysis was performed in two-steps: with and without genes harboring private variants. Enriched functions, such as common gene ontology, biological pathways, shared transcription factor or miRNA binding sites, were reported using the default g:SCS method for significance threshold determination. It is worth mentioning that significant enrichment of protein-protein interactions, derived from the BioGRID database [79], does not imply that all genes with significant enrichment p-value are interacting with each other, but simply indicates which query genes are present in the entire BioGRID dataset. The actual number of interactions and associated genes can be visualized from the network output. The enriched annotations and their gene members were confirmed by literature searches. Furthermore, the SNPshot text-mining tool for PubMed abstracts was used to explore if any of the private homozygous nsSNP-containing genes have clinical or experimental evidence for gene-drug or gene-disease associations [80].

Results and discussion

Sequence-based variant description in 12 genomes

A total of 797 known homozygous non-synonymous substitutions was observed in 575 different genes (**Table 2.1**). The genomes of the four African individuals harbor on average 73 homozygous nsSNPs (range 58–88), while the eight Caucasian genomes have an average of 63 homozygous nsSNPs (range 50–82). The slight excess in African Americans is not statistically significant ($p = 0.18$, 2-tailed t-test). 456 of the genes (79%) have a single homozygous variant in the 12 genomes, but two genes (*HLA-DRB5* and *ANKRD20A4*) have more than 10, detected in at least 6 individuals.

The vast majority of all of the variants have been observed previously in the 1000Genomes project, with just 29 private homozygous nsSNPs observed in 25 different genes, with a range of 0 to 5 per genome. Private variants were found in 10 out of 12 genomes of CHDWB dataset and are listed in **Supplementary Table A.3**. Most are predicted to be neutral, though they affect nucleotides with a range of conservation levels. Almost one-third of the 25 genes have no defined functions, and only a minor proportion of the genes have been previously associated with a disease or trait.

Among the 575 genes whose nsSNPs are homozygous and present in dbSNP build 137, the fractions with known functions, putative functions, and unknown functions are ~47, 20, and 33%, respectively (**Table 2.3**). Almost 10% of the homozygous nsSNPs are found in four highly represented protein groups: transcriptional regulators, keratin-associated proteins, odorant receptors, and zinc finger-containing proteins.

Table 2.3: Functional annotation of all homozygous nsSNP.

Gene groups	Genes		nsSNPs		Most common proteins (#proteins, #nsSNPs)
	#	%	#	%	
Genes with known functions	268	47%	340	43%	Transcriptional regulator proteins (15 proteins, 17 nsSNPs), Keratin-associated proteins (6 proteins, 9 nsSNPs)
Genes with putative functions	118	20%	167	21%	Potential odorant receptors (21 proteins, 27 nsSNPs), Zinc finger-containing proteins, potentially for transcriptional regulations (16 proteins, 22 nsSNPs)
Genes with unknown functions	189	33%	290	36%	-
Total	575	100%	797	100%	

Six programs were used to predict deleterious variants, and three to indicate the level of sequence conservation at the polymorphic site. The results are summarized in **Figure 2.2**, which shows the cumulative number of predicted deleterious (blue) or highly conserved (red) scores for each of the 797 variants. Although 58% of all homozygous missense variants in the 12 genomes alter conserved sites in all three assessments, almost 40% of these (183/463) are predicted to be functionally neutral by all six programs. 45% of the homozygous rare variants are predicted deleterious by at least one program, and 11% (88 variants) by three or more programs.

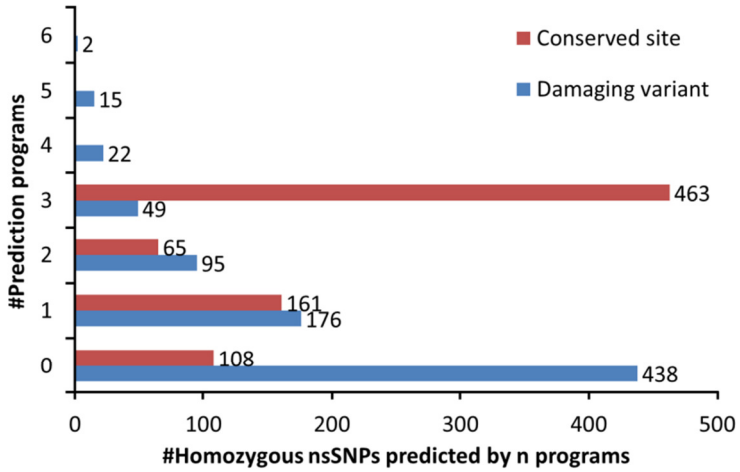


Figure 2.2: Number of deleterious and conserved site predictions. Data labels indicate the numbers of homozygous nsSNPs predicted to be damaging or conserved by n programs (n = 0-6 for deleterious predictions, and n = 0-3 for conservation predictions.)

Differences among the deleteriousness prediction algorithms are underscored by the cumulative score distribution plots in **Figure 2.3**. The solid lines in the top half of each plot are for the 294 homozygous rare variants in the African Americans (black) and 503 homozygous rare variants in the Caucasians (blue). The two curves are not significantly different, and predict that as many as 28% of variants are deleterious (MutationTaster) or as few as 2% (MutationAssessor), with the two most commonly used algorithms, SIFT and PolyPhen2, giving intermediate estimates of 80% neutral. The lower curves show cumulative distributions for a set of 24,703 disease-promoting non-synonymous variants in 1,789 proteins compiled from the MSV3d [43] and SwissVar [44] databases (red dashed curve), as well as from subsets of these disease variants found only in the 23 genes (348 SNPs) harboring homozygous nsSNPs in the four African Americans, or 44 genes (547 SNPs) in the eight Caucasians in our sample (black and blue-dashed curves respectively). All six programs show an elevated tendency to predict known disease-associated variants in the genes harboring homozygous variants in the African Americans

as neutral. This is particularly obvious for the MutationTaster score and least pronounced for SIFT. A similar observation of differences among Asians, Caucasians and Africans in the fraction of damaging SNPs predicted deleterious was made by [81].

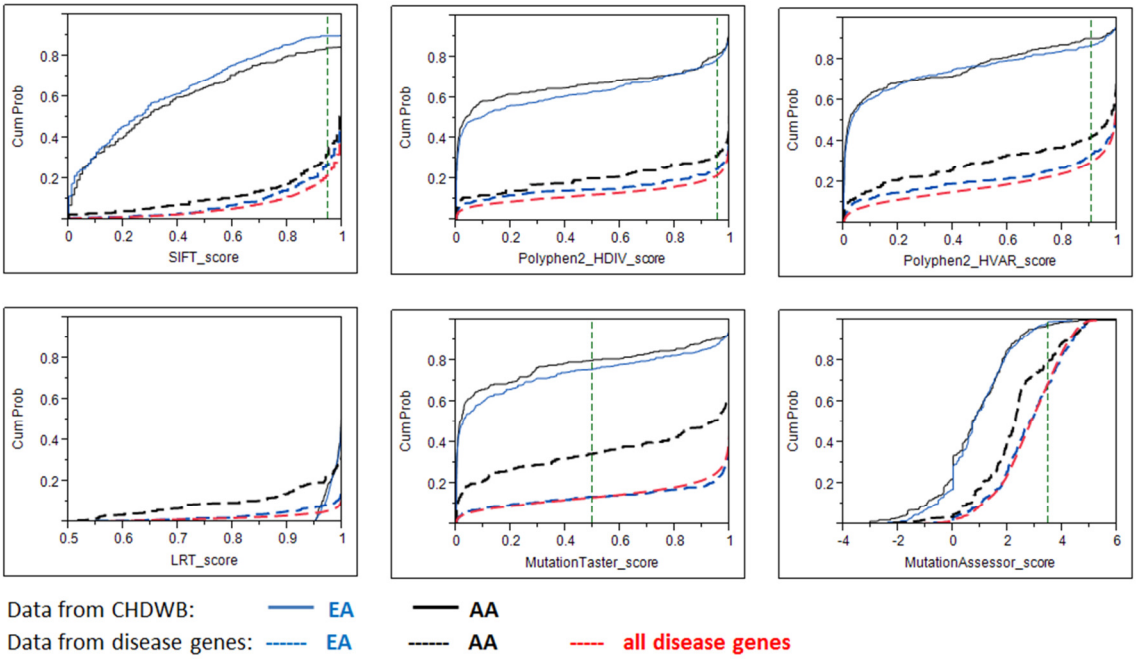


Figure 2.3: Cumulative distribution plots for the six deleterious prediction scores. The X-axis represents the prediction scores, ordered by the deleteriousness. The lowest and the highest scores for each prediction algorithm indicate the neutral and damaging nsSNP, respectively. For each prediction program, the score threshold for defining damaging SNPs is indicated by a vertical dashed green line (threshold for LRT is at 0.999). Five sets of SNP data are shown in each plot. Black solid lines: data from homozygous nsSNPs of four African individuals; blue solid lines: data from homozygous nsSNPs of eight Caucasian individuals; red dashed lines: data from all disease-causing nsSNPs (n = 24,703 nsSNPs in 1,789 proteins); black dashed lines: data from disease-causing nsSNPs found in homozygous nsSNP-containing genes of the four African individuals (n = 348 nsSNPs in 23 genes); blue dashed lines: data from disease-causing nsSNPs found in homozygous nsSNP-containing genes of the eight Caucasian individuals (n = 547 nsSNPs in 44 genes). All disease-causing nsSNPs were retrieved from MSV3d [43] and SwissVar [44].

The notable differences in deleterious predictions given a set of disease-promoting variants found in African American and Caucasian samples (**Figure 2.3**, lower curves)

suggest these prediction algorithms may have population-specific effects. To further investigate that this observation is not an artifact of the small samples, we plotted similar curves for a set of 1,789 genes harboring disease SNPs (total 24,703 SNPs). The genes were classified into four groups depending upon population bias of their SNPs, using %MAF difference cutoff of ± 5% between European American (EA) and African American (AA) populations. The four gene groups are: EA bias, AA bias, EA&AA bias, and no bias. Unlike Figure 3 where homozygous variants in the African Americans were consistently predicted to be more neutral, the larger-sample plots (**Supplementary Figure A.2**) illustrate a small difference in the cumulative scores of population specific-disease SNPs, with the exception of a noticeable prediction bias of MutationTaster. The results highlight the need for development of mutation assessment pipelines that go beyond these algorithms, particularly when evaluating homozygous nsSNPs of non-Caucasian genomes.

Association-adjusted consensus deleterious scheme (AACDS) for variant classification

Consequently, we developed a ranking system that classifies homozygous nsSNPs into eight categories according to the overlap of (i) consensus deleterious prediction, (ii) documentation that the SNP causes a disease, and evidence that the SNP is in a gene that has been associated with (iii) a disease or (iv) a quantitative trait (**Figure 2.1**). Category 1 contains documented disease-causing nsSNPs. Categories 2A and 3A represent nsSNPs in genes that have known associations with diseases (2A) or traits (3A), and these are sub-divided into categories 2B and 3B if they are also predicted to be deleterious by three or more programs. Category 4 comprises nsSNPs that are predicted to be damaging but

lie in genes that have no clinical associations. Conversely, Category 5 nsSNPs are located in genes that have disease or trait associations, but variants are predicted to be neutral. Category 6 represents neutral nsSNPs, whose genes have no clinical relations. **Table 2.2** lists the number of variants in each category from each individual's genome.

The list of disease-promoting nsSNPs was retrieved from MSV3d [43] and SwissVar [44] and is based on manual curation of evidence that the variant is causal, or probably causal, in disease. Most are relatively rare (MAF < 5%) presumed highly penetrant mutations, but an unknown fraction may be false positives. Among the 575 genes in our dataset, 93 harbor disease causal variants (n = 787 nsSNPs). The 797 homozygous nsSNPs in our dataset include 4 category 1 "disease-causing" mutations (1 SNP is present in two individuals) and another 6 listed as probably pathogenic (**Table 2.4**). Another 143 are classified as having unknown effects, leaving 644 polymorphisms presumed not to cause disease in a highly penetrant manner. The 10 probably or known pathogenic missense variants were predicted to be deleterious by 0 to 5 programs (**Table 2.4**).

Table 2.4: List of all four known disease-causal variants and six probable pathogenic variants.

SNP type	Gene	Position, Base change (AA change)	rsID (%MAF EA/AA/All)	Disease [prediction counts]	Grantham score	Stability change	Site annotations
Disease-causal	*ATP6V0A4*	7:138417791 *A-->G (M580T)*	rs3807153* (4.8/18.5/9.4)	Distal renal tubular acidosis (dRTA) with preserved hearing [Del count: 2; Con count: 3]	81	Neutral	TRANSMEM
	MTMR2	11:95569448 *T-->C (N545S)*	rs558018 (0.02/3.9/1.3)	Charcot-Marie-Tooth disease type 4B1 (CMT4B1) [Del count: 2; Con count: 3]	46	Decrease	DOMAIN
	APOE	19:45412079 *C-->T (R176C)*	rs7412 (5.6/8.7/6.6)	Lipoprotein glomerulopathy (LPG) [Del count: 5; Con count: 3]	180	Decrease	REPEAT
	BMP15	X:50658966 *G-->A (A180T)*	rs104894767 (1.4/0.3/1.0)	Premature ovarian failure type 4 (POF4) [Del count: 0; Con count: 1]	58	Neutral	PROPEP
Probable pathogenic	*FRZB*	2:183699584 *G-->C (R324G)*	rs7775 (8.8/28.4/15.4)	Osteoarthritis type 1 (OS1) [Del count: 1; Con count: 3]	125	Neutral	-
	HABP2	10:115348046 *G-->A (G534E)*	rs7080536 (3.9/0.7/2.8)	[Del count: 5; Con count: 3]	98	Decrease	DOMAIN
	HNF1A	12:121416650 *A-->C (I27L)*	rs1169288 (33.5/12.1/26.2)	Insulin-dependent diabetes mellitus type 20 (IDDM20) [Del count: 1; Con count: 3]	5	Neutral	REGION (Dimerization)
	XYLT1	16:17564311 *C-->A (A115S)*	rs61758388 (-/-/1.7)	[Del count: 0; Con count: 3]	99	Neutral	TOPO_DOM
	CYP2A6	19:41354533 *A-->T (L160H)*	rs1801272 (2.5/0.5/1.8)	[Del count: 1; Con count: 3]	99	Decrease	-
	ADA	20:43255220 *T-->C (K80R)*	rs11555566 (6.3/6.8/6.5)	Severe combined immunodeficiency autosomal recessive T-cell-negative/B-cell- negative/NK-cell-negative due to adenosine deaminase deficiency (ADASCID) [Del count: 1; Con count: 3]	26	Decrease	-

*The first SNP (rs3807153) was observed in two individuals.

Table 2.4 (continued)

The minor allele frequency (MAF) in percent listed in the order of European American (EA), African American (AA), and all populations (All), delimited by "/". Prediction counts indicate the number of deleterious predictions (Del count) and conservation predictions (Con count) by six and three programs, respectively. Grantham score determines the similarity in amino acid changes: small (score < 60), intermediate (score 60–99), and large (score ≥ 100). Tertiary classification (increase, decrease, neutral) for protein stability change caused by a SNP was obtained from I-Mutant 2.0 [68], available from MSV3d [43]. Site annotations list any structurally/functionally important sites (molecule processing sites, binding sites, modification sites, etc.) where the altered amino acid residue resides. The information was retrieved from UniProt sequence feature records [51].

Among the 93 disease-associated genes with homozygous nsSNPs in our CHDWB genomes, 18 have homozygous variants present in more than one individual, and these account for 40 nsSNPs, observed at 33 different sites. Only one of these sites, *G56R* in the MYH6 myosin heavy chain, is predicted to be deleterious. Since it is also associated with resting heart rate, it is classified in both categories 2B and 3B. Data for the variant categorization in these 18 disease-associated genes is summarized in **Supplementary Table A.4**. Similarly, we also observed 101 genes with trait associations, including 126 SNPs. Since there are a total of 14 and 21 variants in categories 2B and 3B respectively, most of these cases are restricted to a single individual in the sample of 12. Details for these predicted deleterious variants are listed in **Supplementary Tables A.5** and **A.6**, respectively. On average, each individual carries 3.33 category 1, 2B or 3B homozygous variants (range 0 to 7), and although the two individuals with no variants of this type are both African Americans, there is no significant difference in prevalence relative to the Caucasians ($p = 0.21$, 2-tailed t-test). All of the remaining predicted deleterious variant that do not have disease or trait associations (namely, category 4) are listed in **Supplementary Table A.7**.

The set of 35 disease- or trait-associated SNPs that are also predicted to be deleterious is seven times larger than the set of 5 category 1 "known to be deleterious" mutations. They represent 15% of the 93 disease-associated and 136 trait-associated category 2A and 3A SNPs. **Supplementary Figure A.3** compares the allele frequency distributions of the 2B/3B SNPs relative to all 2A/3A SNPs and shows a tendency to reduced allele frequency, also consistent with them having deleterious effects on fitness. Another 170 homozygous nsSNPs lie in genes that have been associated with diseases or traits but are

not predicted to be deleterious (category 5). Their frequency distribution is approximately equivalent to those of the category 2A and 3A SNPs. The vast majority (539) of the 797 SNPs we have considered are in category 6 and represent the subsets that are least likely to be damaging.

A limitation of our analysis is the uncertainty in the accuracy of phenotypic annotations of SNPs, as well as the variable confidence level in annotations of causal SNPs. We obtained the list of disease-promoting nsSNPs from MSV3d [43] and SwissVar [44]. Most of the variants were classified as either causal variants (to a specific disease), or as polymorphisms. In MSV3d, many variants have ambiguous annotations, e.g., probable-pathogenic, or unknown. In SwissVar, some proteins are noted to have associations with diseases, but the list of variants is not provided.

Supervised structure-based variant evaluation

Personal genome studies indicate that each healthy individual carries a large number of rare homozygous genetic variants [1-4]. While these variants can be found in any structural regions along the genome and can have diverse effects on biological function, Cooper (2010) estimated that as many as 60% of known disease-causing mutations are nsSNPs [5]. This viewpoint simply reflects the more obvious impact of nsSNPs on coding regions than regulatory regions: the substitutions tend to alter the amino acid sequences of the proteins. Amino acid changes are thought to have profound effects on the protein, impacting their structure or function. Furthermore, it is believed that there exist some structure-function relationships for each individual amino acid residue within a protein chain, and the 3D structure is an ideal resource for investigating this

information [82]. Therefore, many algorithms have been developed to assess the effects of amino acid changes within the context of protein 3D structures. Numerous structural features have been used to explain/quantify the changes [83]. A few successful implementations have demonstrated that protein 3D structures add to prediction accuracy [35, 39]. In addition, *in silico* analysis of 3D structures can facilitate variant prioritization because it provides systematic screening of nsSNP effects in the context of the protein structure and suggests which mutations may critically alter the function of the protein.

Implementation of the AACDS classification scheme reduces the number of potentially deleterious variants to a number per genome that can feasibly be evaluated manually on a case-by-case basis. For this purpose, we have devised a further pipeline that involves sequence annotations, extracting either X-ray crystal/NMR structures or homology models from structure databases, and computing a series of predictions that capture protein features. In this way, each of the up to 5 variants in categories 1, 2B or 3B can be assessed in the context of the actual protein. While this approach requires that an individual with experience in protein structures be engaged in the personal genome evaluation, the potential gain in accuracy is likely to be meaningful.

Our preliminary analysis utilized sequence features for all amino acid residues in each protein, obtained from UniProt features records [51]. The entries had been curated and are predicted (and compatible with the protein function), experimentally proven, or determined by resolution of the protein structure. The analysis was restricted to 62% of the nsSNPs, since the remaining fractions do not currently have feature information. **Figure 2.4** illustrates that although the annotated 494 nsSNPs are found in various sequence regions, they are predominately present in transmembrane and protein domains.

These proportions are approximately equivalent to the proportions of each of the 9 types of annotated protein region for all residues in the included proteins (**Supplementary Figure A.4**). Locations of homozygous variants relative to the length of each sequence feature indicate the variants are located throughout the entire sequence length (**Supplementary Figure A.5**).

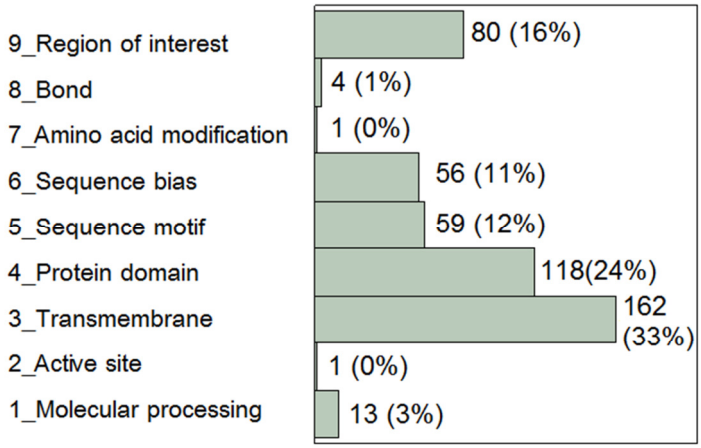


Figure 2.4: Distribution of homozygous nsSNPs by sequence type. Data labels indicate the number (and percentage) of SNPs altering protein residues with each specified sequence features. All sequence features were obtained from the UniProt database [51].

Subsequent detailed analyses involved manual inspection and evaluation of individual proteins. The remainder of this section discusses detailed structural evaluations of two known causative variants, two predicted deleterious variants in proteins that have been associated with a disease or trait, and two predicted deleterious variants for which the clinical associations are inconclusive. The first disease causing nsSNP is the well-known Arginine to Cysteine substitution at residue 176 (residue 158 if omitting the signal peptide) that defines the *APOE2* allele of Apolipoprotein 2 (SNP category 1/2B/3B). This allele has a major influence on lipid transport and is a protective factor against late-onset

Alzheimer's disease and coronary artery disease [84-86]. Homozygosity for *R176C* is also associated with Type III hyperlipoproteinemia (HLPP3) in approximately 2% of cases (though 94% of HLPP3 cases have the genotype) [87]. Onset of the disorder is usually only after menopause in women and rarely manifests before the third decade in men. Several other rare variants in the gene have been annotated to disease, most of which affect intra- and inter-helical salt bridges (**Figure 2.5A**). With the neutral cysteine at position 176 in APOE2 protein, this pattern of salt bridge is eliminated. Although *Arg176* does not interact with the LDL receptor, the *R176C* substitution has been shown to indirectly reduce the receptor-binding activity of APOE [88]. Stability prediction indicates this mutation has neutral effect to the protein stability ($\Delta\Delta G$ = -0.46 kcal mol^{-1}).

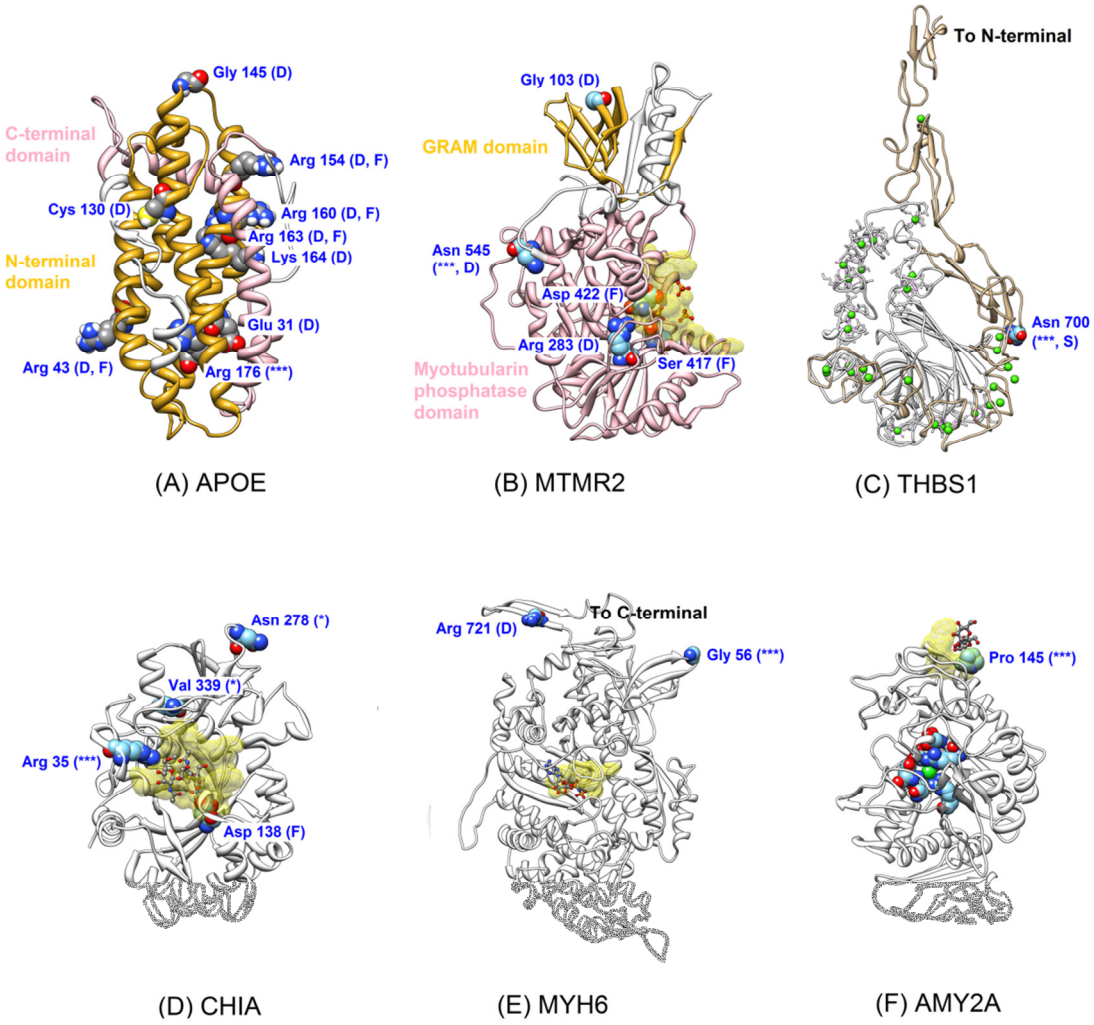


Figure 2.5: Location of variants in six protein structures. A-B describe two causative variants, **C-D** demonstrate two predicted deleterious variants, **E-F** illustrate two predicted deleterious variants whose clinical associations are inconclusive. For all figures, the representations are as follows: ribbon for proteins, ball and stick for ligands, mesh for ligand binding sites, and sphere for amino acid variants. Amino acid variants caused by homozygous or heterozygous nsSNPs are indicated as (***) or (*), respectively. Additional variants that are known to be associated with diseases (D), or affect protein functionality (F), and/or stability (S) are also identified. **A**: Apolipoprotein E (PDB:2L7B). **B**: Myotubularin-related protein 2 (PDB:1LW3). **C**: Thrombospondin-1 (PDB:1UX6 and homology model). Residues 1–548 are missing from the structure, residues 549–829 (brown ribbons) are modeled from human THBS2 (PDB:1YO8), residues 848–1169 (white ribbons) are from a crystal structure of THBS1 (PDB:1UX6). Coordination complexes of amino acid side chains and Ca2+ ions (green spheres) as seen in the crystal structures are indicated with purple lines. The Ca2+ ions with unidentified coordination complexes are derived from the superposition of Ca2+ ions in 1YO8 onto the homology model. **D**: Acidic mammalian chitinase (PDB:3FY1). Three homozygous

nsSNPs were also identified from the same genome, but only two are present in the crystal structure. **E**: Myosin heavy chain 6 (αMyHC) modeled from Myosin heavy chain 7 (βMyHC) (PDB:4DB1). The C-terminal (residues 778–1939) is missing from the template crystal structure. **F**: Pancreatic alpha-amylase (AMY2A) (PDB:3OLE). The Pro145 is located at the end of the extended β-loop and is part of a binding site for α-D-glucose. For clarity, only one α-D-glucose binding site is shown. Other variants known to affect the enzymatic activities are located around the chloride ion (green sphere) in the central vicinity of the protein.

The second example of causative mutation is in Myotubularin-related protein 2 (MTMR2), a putative tyrosine kinase that is associated with Charcot-Marie-Tooth disease type 4B (CMT4B) [89]. One African individual has an Asparagine to a Serine substitution at the position 545 of the protein that has previously been reported as a rare variant in patients with CMT disease [90]. This SNP is classified as category 1. The minor allele frequency of this nsSNP is reported as 3.88% and 0.02% in African Americans, and European Americans, respectively, so the penetrance is much reduced in African Americans since disease prevalence of all forms of CMT is just one in 2,500 [91]. The variant is situated in a conserved site, but only two algorithms predict it to be deleterious. The crystal structure places *Asn545* in a protein domain, but it is not in close proximity with two other causative variants or part of a binding site. This mutation is also predicted to be neutral ($\Delta\Delta G = 0.36$ kcal mol^{-1}) (**Figure 2.5B**).

One predicted damaging nsSNP was found in Thrombospondin 1 (THBS1), a glycoprotein that stabilizes fibrinogen platelet cross-bridges [92]. The homology model indicates the substitution of a Serine for Asparagine at residue 700 of THBS1 occurs at a critical position in one of the calcium-binding domains (green spheres project coordination of CA^{2+} ions) (**Figure 2.5C**). This *Asn700Ser* substitution in THBS1 (SNP category 4) has a prevalence of 8-10% in Europeans [93] and is associated with the

occurrence of premature (age < 45) coronary heart disease in both homozygous and heterozygous individuals [92]. However, a study of *Asn702Ser* in THBS2 (homologous to 700 in THBS1) demonstrated that this variant does not alter calcium-binding capability. Instead, it induces a local conformational change leading to destabilization of surrounding structures [93], consistent with the computational prediction that the variant has a destabilizing $\Delta\Delta G$ of 0.58 kcal mol^{-1}.

The second example of a predicted deleterious nsSNP is a missense substitution in Acidic mammalian chitinase (CHIA), an enzyme that stimulates chemokine production by pulmonary epithelial cells. *Arg35Trp* (SNP category 4) is located in a buried site, where it causes changes in residue side chain volume, charge, polarity, and hydrophobicity (**Figure 2.5D**). The substitution was predicted to disrupt the hydrophobicity of the protein and increase solvent accessibility of the residue [43]. The individual who is homozygous for this variant also carries three heterozygous nsSNPs (*N278D*, *I339V*, and *V432G*). The first two of these replacements are parts of disulfide bonds, while the third substitution resides in the chitin-binding domain. Interestingly, the *Arg35Trp* mutation is predicted to stabilize the protein ($\Delta\Delta G = -1.19$ kcal mol^{-1}), a finding that may appear counter-intuitive. However, it is suggested that protein flexibility is crucial for enzyme catalysis [66]. The increase in protein stability and the dramatic change in physicochemical properties caused by this homozygous nsSNP, along with the disulfide bond reduction from heterozygous variants, strongly suggest the possibility for protein malfunction in this individual. As far as we are aware, he does not have asthma or an aberrant T-helper mediated inflammatory response, but deeper clinical investigation may be warranted.

The last two examples highlight cases where the variant is predicted deleterious but its clinical associations are inconclusive. The first example is a Glycine to Arginine substitution at residue 56 of Myosin heavy chain alpha (MYH6) (SNP category 2B/3B). As mentioned earlier, this is the only predicted deleterious variant among a set of 18 disease-associated genes with variants present in more than one individual (**Supplementary Table A.4**). SNP databases indicate six well known causative variants in this gene that lead to familial hypertrophic cardiomyopathy and atrial septal defect. Although *G56R* is not one of them, this variant had been previously identified in affected individuals but it does not segregate perfectly with the disease in families of probands [94, 95]. With regard to the homology model (**Figure 2.5E**), many of the known variants associating with heart disease are located in the coiled-coil regions of this protein (missing from the crystal structure) and are not part of the ATPase catalytic site or actin binding site. Nonetheless, the *G56R* found in one of the CHDWB participants is particularly interesting, since the mutation occurs in a myosin head-like domain, a key component for muscle contraction, and with a large degree of amino acid change (Grantham score = 125). Stability prediction also suggests this variant destabilizes the protein ($\Delta\Delta G = 1.10$ kcal mol^{-1}).

The second example of a variant with uncertain functional effect is taken from a set of 143 SNPs (18% of CHDWB dataset) that do not currently have phenotypic annotations. Among these, 117 SNPs have neither disease nor trait association at the gene level. Form 143 SNPs, we identified 15 variants predicted to be damaging, of which 8 are located in genes with no clinical associations. Our example is a Proline to Serine substitution at residue 145 of Pancreatic alpha-amylase (AMY2A) (SNP category 4). In addition to one

calcium ion and one chloride ion per subunit, the protein is able to bind to several ligands throughout the structure. Mutagenesis studies identified several amino acid residues that are essential for the protein's catalytic activity and affinity to calcium and chloride ions, but the impact of *Pro145Ser* has not been established [96, 97]. Despite the limited information of the variant, crystal structure indicates *Pro145* is part of one, among many, binding sites for alpha-D-glucose (**Figure 2.5F**). In general, amino acids with similar physicochemical properties may substitute each other while maintaining the protein's functionality. One study demonstrates some uncommonly predominant inter-species amino acid variations, such as serine-proline pairs or glutamic acid-alanine pairs [98]. The notable feature corresponds with the proline to serine substitution caused by this SNP. It is well known that the proline residue is sterically restricted and that it tends to disrupt regular secondary structural elements. Most proline residues are found in very tight turns or on protein surface [99]. The unusual occurrence of *Pro145*, especially in the extended β-loop indicates that this residue is essential for proper protein folding. Further investigation revealed that this residue is in the *cis* isomer, a very rare phenomenon since proline residues are exclusively synthesized as the *trans* isomer. In fact, AMY2A contains two *cis*-proline residues (*Pro69* and *Pro145*); both are located in the loop regions. It had been suggested that the two residues help accommodate a sharp turn of the β-loop [100]. Substitution of proline to serine is predicted to be highly stabilizing for residue 145 ($\Delta\Delta G = -2.77$ kcal mol^{-1}), but highly destabilizing for residue 69 ($\Delta\Delta G =$ 5.23 kcal mol^{-1}). In any case, strong stability changes are suggested to cause protein malfunction and may lead to disease(s) [66].

Automated structure-based variant evaluation

To facilitate high throughput evaluation of protein structures, we devised a structural analysis pipeline that assesses the functionality of protein residues using data directly obtained from the atomic coordinates or from computational predictions. Using this approach, a list of potentially deleterious variants from a structural perspective can be generated rapidly, providing a way to integrate structural analysis into the variant categorization scheme.

The four areas of automated structure-based variant analysis include stability, flexibility, and potential to disrupt protein-protein or protein-small molecule interactions. Many mutations disrupt these structural features and as a result, lead to altered protein functions or diseases. Our assumption is that the analysis may be able to identify some variants with strong effects. The results are summarized in **Supplementary Table A.8**, which indicates that predicted deleterious variants show a wide variety of residue features. In general, SNPs that do not cause stability change ($\Delta\Delta G < \pm 0.5$ kcal mol^{-1}) tend to be non-deleterious, but not *vice versa*. Four amino acid residues have B-factors of Cα atoms larger than 60 $Å^2$, a characteristic which may indicate that the atom is disordered due to dynamic motion, or may be an artifact of crystal imperfection. However, three out of these four amino acids are predicted to be at rigid sites and do not induce conformational switches. Another three residues are predicted to lie within a cluster of conserved residues on a protein surface, but none is categorized as damaging variants. Lastly, we found only one amino acid mutation that is located within the binding site, but the variant is in categories 2A and so not predicted to be damaging.

Imposing the constraint that the 3D structures must be of high quality, our initial analysis was restricted to only 24 experimentally-determined structures and 1 high quality theoretical model. Their sequence coverage ranges from 23–100% (average 69%) (**Supplementary Table A.2**). However, the implementation can be further applied to any available structures. For example, using the automated Phyre2 homology modeling server with single/multiple template methodology [101], we were able to model an additional 77 full length proteins with high confidence. Each full length protein model has a percentage of residues modeled at >90% confidence in the range of 59-99% (average 86%). For larger proteins (>1,300 amino acids), we truncated them into smaller domain(s) and our modeling attempt returned 10 models with confidence between 96-100%.

Enrichment for mutations disrupting protein interactions

As a parallel approach to evaluating the deleterious potential in homozygous protein substitutions, we used g:Cocoa [78] to evaluate whether there is an enrichment for proteins that have similar functions. The analysis revealed four significant gene annotations that include a significant number of the queried genes from more than one individual. These four terms are: X-linked recessive inheritance, epithelial cell signaling in *H.pylori* infection, microRNA miR-708 binding sites, and Rho GTPase signaling pathway (**Supplementary Figure A.6**).

Although 96 homozygous nsSNPs were identified on the X chromosome, only 1 nsSNP has been documented as a causal variant (X:50658966 $G \rightarrow A$) (**Table 2.4**). Genes involved in X-linked recessive inheritance from the 12 genome data were identified in one African female and four Caucasian males (**Supplementary Table A.9**). We

identified only two predicted damaging SNPs. One was found in *TBX22* that is associated with X-linked cleft palate. The other is located in *SYTL5*, a trait gene associated with erectile dysfunction and prostate cancer treatment. The remaining X chromosome variants are predicted as neutral. Two male individuals have the same mutation in the *F9* gene, for which reduced function can result in hemophilia B (HEMB). Unlike hemophilia A, symptoms of HEMB are usually milder or can be asymptomatic. Three male individuals also carry an identical SNP in *FRMD7* gene. Malfunctions of this gene can cause nystagmus congenital X- linked type 1 (NYS1), a condition that appears at birth and up to three months old. The indications are spontaneous and involuntary ocular oscillations. Given that most X-chromosome variants are predicted neutral and there is no indication that either of these individuals have these conditions, it is unlikely that the associated disease or trait will develop.

Finally, we used BioGRID [79] to evaluate enrichment for proteins that form physical interaction networks, and found that one of the 12 individual's genomes has 7 mutations potentially involved in an unusually high number of contacts (**Supplementary Figure A.7**). Three of these proteins (FHL2, STK17A, and DSP) are linked together by other interacting partners. The *FHL2* gene encodes a four and a half LIM domain protein that acts as a molecular transmitter between signaling pathways and transcriptional regulation. The wild type amino acid affected by the homozygous missense variant of the *FHL2* gene is evolutionarily conserved, but the *Arg177Gln* substitution is only predicted to be deleterious by one program. This variant is not predicted to affect protein stability and the protein is not associated with a disease or trait. However, we mention it as an example of how this approach may highlight networks of proteins, where subtle modification of

multiple partners may result in cumulative disruption that would lead to disease under a multiplicative burden of rare variants model.

Conclusions

Intensive resource requirements and the associated costs make it infeasible to experimentally verify the effect of every genetic variation. At this stage of human genome study, *in silico* predictions play an important role in identifying putative functional variants. While a sequence-based approach is the current standard practice for assessing SNP effects, there are still some concerns that sequence conservation alone is not a reliable predictor of deleteriousness. In this study, we propose the AACDS classification scheme using variant annotation and sequence-based predictions. We used AACDS to classify homozygous nsSNPs found in the genomes of twelve healthy individuals into eight categories according to the consensus sequence-based deleterious prediction, types of mutation (disease-associated vs. neutral), and information on disease- or trait-associations with the gene. The classification scheme provides a comprehensive framework for prioritizing a list of SNPs suitable for detailed evaluation, in this study reducing the evaluation space from 826 to 98 variants (in categories 1, 2B, 3B, and 4). An online tool for computing the AACDS scores for any variant is provided at http://www.cig.gatech.edu/tools.

Several previous studies have shown that structural information plays an important role in understanding the relationship between genetic variation and the structure and function of the protein. The addition of 3D structural analysis following AACDS classification demonstrates how structure data can complement sequence-based prediction, and

highlights how functional interpretation can in some cases be inferred exclusively from 3D structures. By using a combination of solved structures or high quality homology models for all human proteins, we demonstrate that up to 117 of the 575 proteins bearing homozygous mutations in our CHDWB dataset are available for detailed SNP evaluation, providing detailed analysis of the 150 prioritized variants.

Acknowledgements

This work was supported by start-up funds from the Georgia Tech Research Foundation to GG, and TP was supported by the School of Biology at Georgia Tech.

CHAPTER 3: AACDS—A DATABASE FOR PERSONAL GENOME INTERPRETATION [102]

Abstract

Background: Incorporation of diverse data sources add value to genomic studies, especially for annotation and categorization of personal genome variation, based on the putative functionality of variants. The database for Association-Adjusted Consensus Deleterious Scheme (AACDS) and its web application deliver a novel approach to assess genetic variation; the schema combines commonly used conservation-based measures of deleteriousness with phenotypic and/or disease association statistics to prioritize functional assessments.

Description: The AACDS database covers over 68 million nsSNPs in approximately 18,000 human genes. The simple but interrelated queries classify each variant into an 8-level category, according to its consensus deleterious prediction and the presence/absence of clinical or phenotypic association data. Retrieval of AACDS classes can be performed through a simple search platform. Given a list of single nucleotide variants located at chromosomal locations, or within gene or protein sequences, the AACDS web application returns the AACDS category for each variant, along with known data. The categories can be ranked, enabling straight-forward interpretation of relative likelihood of functionality. The ranking thus facilitates improved efficiency in prioritizing further detailed evaluation of key variants within a personal genome.

Conclusions: The AACDS database is built upon integrated knowledge of variant data, with the aim of relating clinical phenotypes to predictions of variant deleteriousness. The schema highlights a list of variants in individual genomes that are worth examining. The AACDS web application is available at http://cig.gatech.edu/tools.

Background

Non-synonymous single nucleotide polymorphism (nsSNP) is one of the most common forms of genomic variability. About 60% of known disease-causing mutations are nsSNPs [5]. One of the major goals for personal genomics is to identify a subset of variants that have the potential to influence an individual's health. Each individual genome is estimated to contain roughly ten thousand nsSNPs [41, 46, 103]. The assessment of deleteriousness for SNPs is commonly performed on a per variant basis, by using many available computational tools that typically classify each SNP into two groups: benign and damaging. Although many prediction programs have been proven to have acceptable accuracy, mostly in the range of 70-80% [22], it is deemed an advantage to incorporate more data into the assessment [28].

In our recent study on interpretation of personal genome data [40], we developed "An association-adjusted consensus deleterious scheme" (AACDS) to facilitate variant prioritization of personal genome studies. AACDS is constructed from the combination of existing databases that implicate the variant in disease or a phenotype, and traditional sequence-based predictions. It assigns a variant into an 8-level category. Not only does AACDS incorporate the clinical or phenotypic annotations of the genomic variants in an

individual, it also narrows down the variants to a subset that is appropriate for further follow-up experiments and validation with respect to individualized health profiles.

To promote the utility of our variant classification schema AACDS, we have implemented the assessments in a database-driven web application that allows users to search the AACDS categories and relevant information for a list of user-defined variants. The AACDS website aims to provide a user-friendly platform for anyone who is interested in personal genome interpretation. The database schema was designed to cover the annotated list of functional variants (31,092 disease-associated amino acid variants in 3,363 genes), 4,225 pairs of gene-disease associations, 5,113 pairs of gene-trait associations, and all possible coding genomic variants in 18,349 human genes (n= 68,165,196 nsSNPs). Therefore, our newly developed database-driven web application for AACDS can serve as a tool to generate the best estimate of clinical significance of each variant from the large and growing accumulation of personal genome data. In addition to identifying causal variants or variants in disease- or trait- associated genes from a list of genomic variability, the application also allows further functional analyses of all SNPs in any gene of interest.

Although many tools and databases exist for the purpose of variant prioritization and/or personal genome interpretation, we are not aware of any tool with similar features as ours, especially in the categorization of genomic variants. Our AACDS tool allows SNP evaluations to be performed simultaneously on the ground of deleterious predictions, direct connections between variants to diseases, and associated traits and diseases to the genes. The tool assigns an AACDS class to each individual SNP; it also reports the overall AACDS statistics for a given genome. The classification and the ranking of SNPs

are particularly significant and original since it assists effortless interpretations of whole genome SNP searches. The results facilitate the identification of high impact variants within a genome in an effective and efficient manner.

Compared to aggregative variant association methods such as in VAAST 2.0 [104], our tool does not require that users have prior knowledge of various additional genomic attributes to perform the search and interpret the results. VAAST requires not only target and background genome datasets, but also the user's defined set of genes and prior knowledge of genetic parameters (e.g. inheritance, penetrance, locus heterogeneity, allele frequency, etc.) in order to search for causal SNPs or genes [104]. The search pipeline is not designed for evaluation of all genomic variants nor to be used as a simple look-up utility.

Two recent genome analysis tools, eXtasy [105] and Phen-Gen (www.phen-gen.com/) [106], introduce a new phase of genome interpretation in which the tools link genome variants to a specific phenotype. Although both tools have great potential for guiding diagnostics of rare disorders through the identification of phenotype-specific causal variants, the evaluations are performed on a per disease basis. Most personal genome variants are likely to be neutral and contain a minimal number of annotated disease SNPs [40, 42]; the individuals are considerably healthy or are unlikely to have noticeable clinical phenotypes [46]. These limitations represent a significant challenge for personal genome variant annotation for sub-clinical phenotypes, which AACDS is designed to help in the interpretation of.

Construction and content

The AACDS site (http://cig.gatech.edu/tools) serves as an interface for queries of the AACDS databases, which is built to categorize nsSNPs into an 8-level class, based on the consensus predicted deleteriousness and the evidences of disease or complex trait associations with a SNP of its gene. The AACDS database includes a set of 68,165,196 nsSNPs that can be found in a human genome. The AACDS website allows users to retrieve the AACDS classification and relevant information of variants in genes of interest.

Data sources

To facilitate the variant mapping of various data formats (chromosome coordinates, gene names, protein names), we chose UniProt [51] as the main database to interrelate with the others. UniProt accession numbers represent the unique identifiers for gene products. The unique accessions allow a direct lookup of the disease-association data from the selected SNP databases: MSV3d [43] and SwissVar [44]. A list of 20,277 reviewed human proteins was compiled from UniProt [51] (accessed 11/01/2013). This protein set represents the gene products of 19,700 genes. (Some genes have multiple UniProt accession numbers due to the presence of multiple protein isoforms, thus explaining the differences in the absolute numbers of UniProt accessions and genes.)

Next, we used dbNSFP v2.1 [15] (released 10/03/2013) to extract all possible locations of SNPs within each gene. The database provides the translation of nucleotide variants into alternate amino acids. Amino acid variations were indexed with respects to the correspondent proteins. All functional predictions (benign vs. damaging) of a SNP were

retrieved from the pre-computed scores for six sequence-based deleterious predictors available from dbNSFP v2.1 [15]. To resolve discrepancies among prediction algorithms, we assigned the level of deleteriousness using the consensus prediction. A variant is regarded as "deleterious" if ≥ 3/6 predictors reported the variant as "deleterious" and as "non-deleterious" if the predictions suggest otherwise. Later, the initial set of SNPs had been filtered such that our AACDS database excludes all variants that do not currently have all of the six pre-computed deleteriousness scores available and variants in chromosome locations where only transcript IDs are available but cannot be mapped to a single gene. The high confidence dataset includes a total of 68,165,196 nsSNP locations in 18,349 genes (18,390 gene products).

Gene-trait associations were retrieved from the NHGRI genome wide association studies (GWAS) catalog [107], available from dbNSFP v2.1 [15]. Additional information provided at the AACDS website includes essential information of each variant; i.e., dbSNP reference SNP ID number (db138 release, downloaded from the NCBI's FTP site at ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/BED/, accessed 1/16/2014), gene name and protein name from UniProt database (accessed 11/01/2013) [51]), and population-specific minor allele frequencies (retrieved from dbNSFP v2.1 [15]).

Database construction

AACDS was designed as a relational database on a MySQL server. The data relationships are presented in **Figure 3.1**. In-house Perl scripts were used to extract variant information from the various aforementioned data sources and to prepare the MySQL data structure.

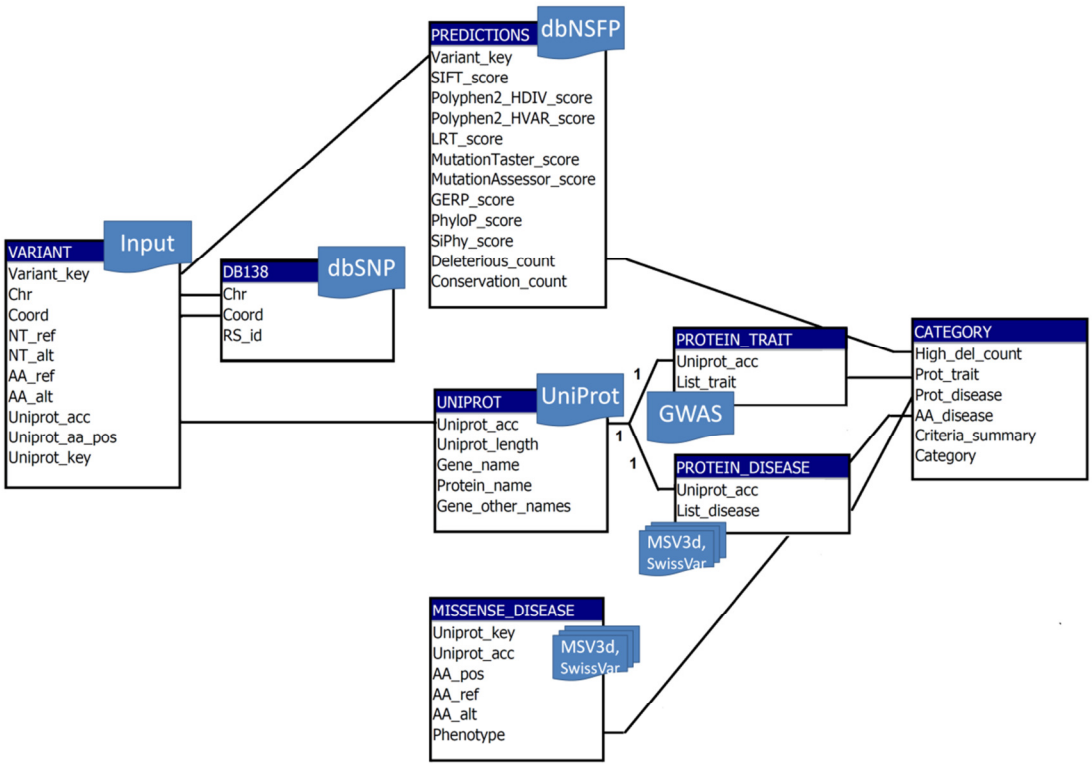


Figure 3.1: AACDS database schema. AACDS database constructs its data relationships from several sources. Each SNP receives the deleterious predictions of six predictors, available from dbNSFP v2.1 [15]. The consensus deleteriousness of the variant is assigned; variants with high deleterious count refer to those which are predicted to be damaging by at least 3 predictors. Documentation on gene-trait associations from GWAS were also obtained from dbNSFP v2.1 [15]. A list of associated diseases to the variant or to its gene is populated from publicly available SNP databases (MSV3d [43] and SwissVar [44]). The AACDS category for each variant is assigned from whether: the variant has a high deleterious count, its gene is associated with a trait, its gene is associated with a disease, or the variant is documented as a disease causal variant.

Our original paper describes the AACDS as an 8-level category (variant categories 1, 2A, 2B, 3A, 3B, 4, 5, and 6) [40]. However, many SNPs cannot be exclusively defined into one class, therefore, a maximum of 12 classes are reported in this implementation to represent all distinct conditions possible when joining multiple assigned AACDS categories together; e.g. a variant can be in class 1/2B, 1/3B, or 1/2B/3B, etc. (**Table 3.1**). **Figure 3.2** illustrates the number of nsSNPs in each of the 12 combined AACDS classes.

Table 3.1: Descriptions of the 12 combined AACDSS classes. The original paper describes the AACDS as an 8-level category (variant categories 1, 2A, 2B, 3A, 3B, 4, 5, and 6), but many variants can belong to multiple classes. Twelve combined AACDSS classes are used in this AACDS database to enable the informative description of the variants.

AACDS classes	Features of SNPs				Descriptions of SNPs
	Diseasing-causing	Predicted deleterious	In disease gene	In trait gene	
1	yes	no	no	no	Disease-causing (but not located in gene with disease- or trait-associations nor predicted as deleterious by most programs)
1, 2B	yes	yes	yes	no	Disease-causing, predicted as deleterious by most programs, located in gene with disease-associations (but no gene-trait associations)
1, 2B, 3B	yes	yes	yes	yes	Disease-causing, predicted as deleterious by most programs, located in gene with disease and trait-associations
1, 5	yes	no	yes	yes	Disease-causing, located in gene with disease- and trait-associations (but most programs predicted it to be benign)
2A	no	no	yes	no	Located in gene with disease-associations (but no other implications)
2B	no	yes	yes	no	Predicted deleterious by most programs, located in gene with disease-associations (but not a causal variant)
2B, 3B	no	yes	yes	yes	Predicted deleterious by most programs, located in gene with disease and trait-associations (but not a causal variant)
3A	no	no	no	yes	Located in gene with trait-associations (but no other implications)
3B	no	yes	no	yes	Predicted as deleterious by most programs, located in gene with trait-associations (but not a causal variant)
4	no	yes	no	no	Predicted as deleterious by most programs (but no other implications)
5	no	no	yes	yes	Located in gene with disease and trait-associations (but not a causal variant nor predicted as deleterious)
6	no	no	no	no	No implications

Column descriptions: (1) disease-causing: if MSV3d [43] and/or SwissVar [44] indicate the variant is a disease-causal; (2) predicted deleterious: ≥ 3/6 programs predict the variant to be deleterious; (3) in disease gene: if MSV3d [43] and/or SwissVar [44] indicate the gene has disease associations; (4) in trait gene: if GWAS [45] indicates the gene has trait associations.

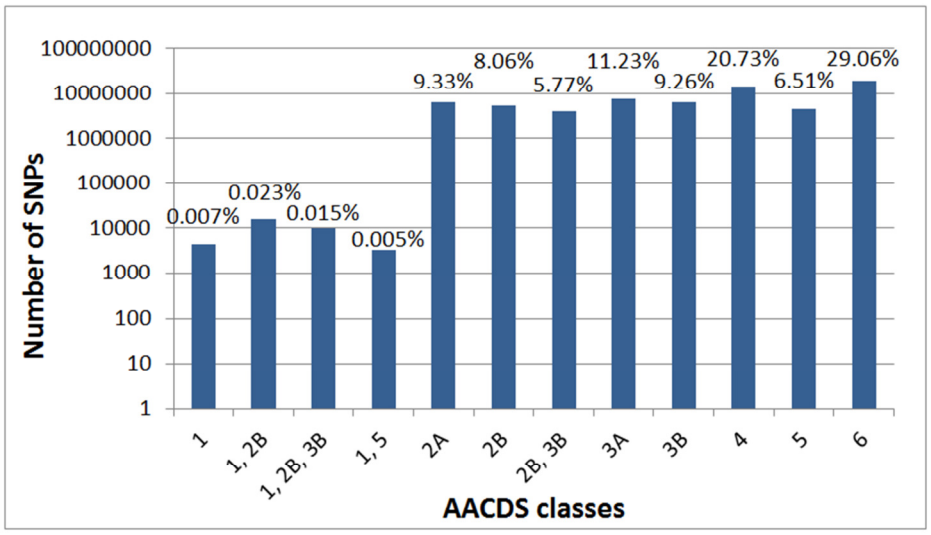


Figure 3.2: Number of nsSNPs within each AACDS category. Each variant is exclusively defined into one of the 12 combined AACDSS classes.

The SNP-disease or gene-disease associations were collected from all associations that have been documented in either of the SNP databases (SwissVar (accessed 11/01/2013) [44] and MSV3d (released 07/29/2012) [43]). It is worth mentioning that the two databases employ different formats for clinical annotations, such as the minor differences in disease names and the sub-categorization of certain diseases. We did not attempt to standardize these terms. Similar association records for a particular SNP or a gene from the two data sources were dealt with by reporting only the record which has the most detailed descriptions. Some SNPs have ambiguous clinical annotations; for example, when one of the two databases documents a SNP as a disease-associated variant, but the other suggests it is a polymorphism or has missing data, the intuition we followed was to regard the variant to have clinical associations. Such example includes 2,797 and 6,738 pairs of associations which have only MSV3d [43] or SwissVar [44] annotations respectively (21,557 variants have annotations from both sources).

In total 31,092 instances of variant-disease associations and 4,225 pairs of gene-disease associations were included in our database. The number of genes whose gene-trait associations were identified from GWAS is 5,113.

To ensure that the search results are returned quickly, we performed the computation of AACDS for all variants and utilized the assigned categories as the pre-computed variant classification during web searching.

The online service of the AACDS database was implemented in PHP, MySQL, JavaScript and Apache. The AACDS website can be accessed at http://cig.gatech.edu/tools. All standard browsers are supported.

Utility and Discussion

Our AACDS web application allows users to retrieve AACDS classifications and the relevant information of variants or variants in genes of interest. **Figure 3.3A** illustrates the three major components of the website: (1) Variant query, (2) Gene query, and (3) AACDS-based genome analysis. Users can search the AACDS database via a single entry query or a batch query. The batch query permits the practical analysis of personal genome data, since users can upload a list of variants of unlimited size and retrieve the results in plain text formats (*.txt* or *.csv*) for external use.

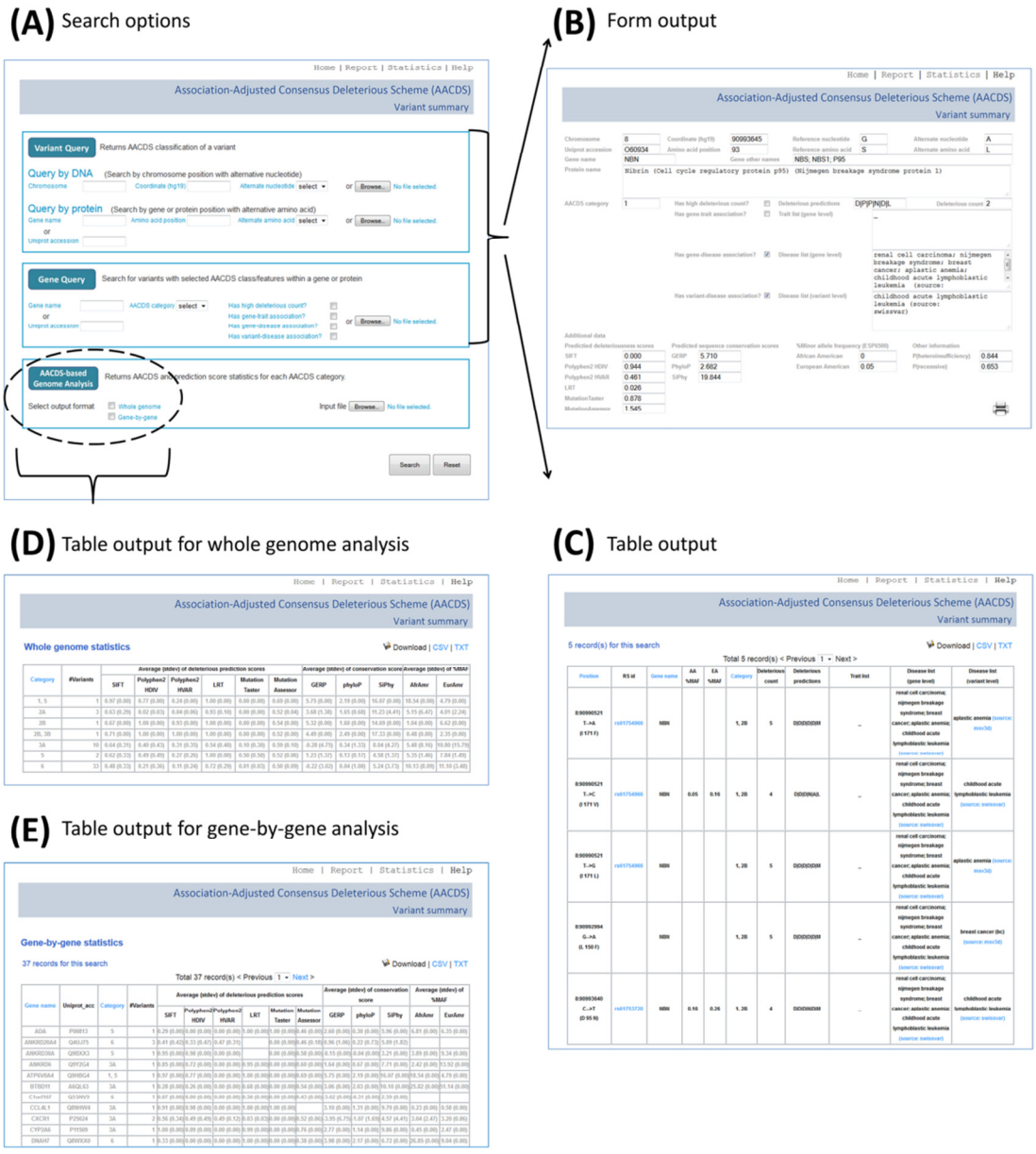


Figure 3.3: Overview of the AACDS web interface. (**A**) The three query options: Variant query, Gene query, and AACDS-based genome analysis. (**B-E**) Example outputs in form and tabular formats. Included in the tabular outputs are the direct links to dbSNP [49] and to the original sources [43, 44] of clinical data.

Performing the search via a single entry query

1) *Variant query*

Users can search for the AACDS classification of their variant of interest, by providing some search parameters. For query by DNA: chromosome number, hg19 coordinate and alternative nucleotide. For query by protein: gene name or Uniprot accession number, amino acid position, and alternative amino acid.

The website outputs a variant summary page, which reports the AACDS category of the variant and its relevant information, along with any additional variant data (**Figure 3.3B**). Note that all the deleterious and conservation scores reported here are the original scores used in their corresponding papers; i.e., we did not perform any re-scaling of the scores.

2) *Gene query*

Users can retrieve a list of variants within a gene whose variants' characteristics match the user's interest. If a particular AACDS class is specified, the website returns all SNPs that belong to the searching AACDS category. If any of the four features (high deleterious count, has gene-trait association, has gene-disease-association, and has variant-disease association) are specified, a list of variants whose characteristics are compatible with the searching feature is returned. The SQL "AND" statement is used to extract the list of variants which match multiple search terms; for example, searching for variants that are documented as causal variants whereas most of the deleterious prediction algorithms suggest it is benign, or searching for variants in disease-associated genes that are also predicted to be deleterious by ≥ 3/6 predictors.

When more than one variant meets the search criteria, a summary table (**Figure 3.3C**) is returned in addition to the results in form formats, which display up to the first 25 entries. The displayed table provides a short description (11 attributes) of the variants; users can also download the complete table (37 attributes) through the "download" button. Exported file types include *.csv* and *.txt*.

3) *AACDS-based genome analysis*

For this search feature, we aim to provide the overall statistics for a set of SNPs found in an individual's genome. Users can perform the AACDS-based genome analysis on two levels: (1) Whole genome statistics and (2) Gene-by-gene statistics. **Figures 3.3D** and **3.3E** demonstrate the output examples from the two analyses, respectively. In either case, the schema classifies SNPs into several groups, based on the assigned AACDS classes. The results can be ranked by gene names or by AACDS groups.

In addition to the number of variants within each AACDS class, the tabular output also presents the average (and the standard deviation) for all six deleterious scores, three conservation scores, and two population-specific minor allele frequencies. Note that the raw scores for the first five deleterious predictors were re-scaled to [0, 1] for comparison purposes; a score closer to 1 represents a stronger (deleterious) effect of a variant. A MutationAssessor score of > 3.5 designates high functional impact, hence "deleterious" [18]. The averages of original scores can be obtained via the download link.

Performing the search using batch query

For each of the above analyses, a batch search is possible if users provide a *.txt* file (tab delimited) with the required information in specific formats as described in the website's

help page. An example format for a batch search is shown below; other search types will have slight differences in the required fields. For the example below, column legends are chromosome, coordinate (hg19), reference nucleotide, and alternate nucleotide, respectively.

```
Chr:10      26781257    T     A
Chr:10      26781257    T     C
Chr:10      26781257    T     G
```

Conclusions

The integration of both sequence-based deleterious prediction and clinical association data in our AACDS algorithm provides a novel approach to integrative variant classification for personal genomes. Manual inspection of a variant for both predicted deleteriousness and phenotypic association is possible, but certainly not practical for analyzing large genome data. For this reason, the implementation of a database-driven web application is considered to be an important tool for promoting the utility of the AACDS. We believe that with the scope of our database coverage, both in terms of genomic variations and phenotypic data, this web application will help to bring a comprehensive framework of personal genome interpretation to a more practical level.

We will keep on refining the database so that it offers AACDS classes for the most complete set of SNPs in a human genome. The improvement may include quality control and subsequent addition of variants in the remaining genes once their curated protein sequences are available, the inclusion of clinical and trait associations from other data sources, the update of medical terminology so that they are consistent with standard terms

used by the International Classification of Diseases (ICD) [108], and the implementation of an automatic online update with the selected data sources.

Availability and requirements

AACDS is publicly available at http://cig.gatech.edu/tools. It supports any standard browsers. The current implementation does not have an automatic online update feature, but we will regularly check for new releases of our selected external databases. The update for AACDS will be performed quarterly and upon major releases of MSV3D [43], SwissVar [44], and dbSNP [109].

Acknowledgments

This work was supported by start-up funds from the Georgia Tech Research Foundation to GG, and TP was supported by the School of Biology at Georgia Tech.

CHAPTER 4: SDS, A STRUCTURAL DISRUPTION SCORE FOR ASSESSMENT OF MISSENSE VARIANT DELETERIOUSNESS [110]

Abstract

We have developed a novel structure-based evaluation for missense variants that explicitly models protein structure and amino acid properties to predict the likelihood that a variant disrupts protein function. A structural disruption score (SDS) is introduced as a measure to depict the likelihood that a case variant is functional. The score is constructed using characteristics that distinguish between causal and neutral variants within a group of proteins. The SDS score is correlated with standard sequence-based deleteriousness, but shows promise for improving discrimination between neutral and causal variants at less conserved sites.

The prediction was performed on 3-dimentional structures of 57 gene products whose homozygous SNPs were identified as case-exclusive variants in an exome sequencing study of epilepsy disorders. We contrasted the candidate epilepsy variants with scores for likely benign variants found in the EVS database, and for positive control variants in the same genes that are suspected to promote a range of diseases. To derive a characteristic profile of damaging SNPs, we transformed continuous scores into categorical variables based on the score distribution of each measurement, collected from all possible SNPs in this protein set, where extreme measures were assumed to be deleterious. A second epilepsy dataset was used to replicate the findings.

Causal variants tend to receive higher sequence-based deleterious scores, induce larger physico-chemical changes between amino acid pairs, locate in protein domains, buried sites or on conserved protein surface clusters, and cause protein destabilization, relative to negative controls. These measures were agglomerated for each variant. A list of nine high-priority putative functional variants for epilepsy was generated. Our newly developed SDS protocol facilitates SNP prioritization for experimental validation.

Introduction

Several prediction programs are available to evaluate missense variants as either deleterious (having a strong functional effect) or neutral (having no or only a weak functional effect) from the level of DNA or protein sequence conservation [14]. While existing sequence-based damaging scores agree for the most deleterious variants, predictions for candidate moderate effect variants identified from sequencing studies are not much better than chance. Since there is no clear way to truly evaluate the predictive accuracy of the scores prior to experimental assessment of function, there is scope for development of orthogonal methods for variant prioritization. Our study explores the utility of solely using protein structure-based assessments as a complement to existing sequence-based scores.

Of the commonly used automatic tools for prediction of variant deleteriousness, PolyPhen2 [19] already incorporates protein structure information. It uses an iterative greedy algorithm to select certain features from a restricted training set, and then takes a Bayesian approach to assign each variant into one of four effect categories: probably damaging, possibly damaging, benign, and unknown. However, it does not perform

evaluations on the actual protein structure that each variant is found in. Rather, PolyPhen2 includes experimentally derived-structures that are available for ~10% of the training set. Although the implementation has high accuracy (73-92%) for the identification of true positives in cross-validation data, structural data does not directly contribute to evaluations of novel genes and it is not clear how efficiently the generalized structural characteristic rules used by the algorithm can contrast clinically-associated variants from neutral variants in a diverse gene set.

In this study, we therefore introduce a new approach for assessing the deleteriousness of nonsynonymous single nucleotide polymorphisms (nsSNPs). Our newly developed protocol uses additional information, that is, protein structure-based assessments applied only where the structural solution is available, to complement existing sequence-based scores. More specifically, our evaluation pipeline focuses on functionality of protein residues derived from 3-dimensional (3D) protein structures. We also incorporate multiple classes of structural assessment, namely measures of protein stability, flexibility, protein-protein interaction potential, and small-molecular binding. As several studies [35, 39] have proven that structural information increases classification accuracy of SNPs, we hypothesized that by incorporating results from several structure-based assessments, it may be possible to generate characteristic profiles that enhance prediction of the degree to which a candidate rare variant may disrupt protein function, and lead to disease development.

We applied this newly developed variant assessment protocol to a set of 57 gene products harboring homozygous missense variants, discovered in a recent large-scale exome sequencing study, that are exclusive to epilepsy patients [111]. Epilepsy is a highly

genetically heterogeneous disease, for which each likely causal variant is observed in a small fraction of individuals, likely with variable expressivity and penetrance [112]. As a result, it is difficult to ascertain which variants are truly responsible for the etiology of disease in individual patients. None of the case-exclusive variants documented by Heinzen et al. (2012) had a high enough prevalence to support statistical association with the disease, so experimental tests will be needed to filter putative causal variants. By contrasting the spectrum of structural features of the case variants with positive control known causal variants and negative control neutral variants observed in healthy individuals for the same proteins, we illustrate the potential for structural assessment to prioritize new variants for functionalization.

Materials and Methods

Our analysis pipeline applied sequence- and structure-based assessments to missense mutations and their 3D protein structures to depict the likelihood that a mutation disrupts protein function. Numerous databases and prediction programs were used. The flow diagram of the analysis protocol is illustrated in **Figure 4.1**.

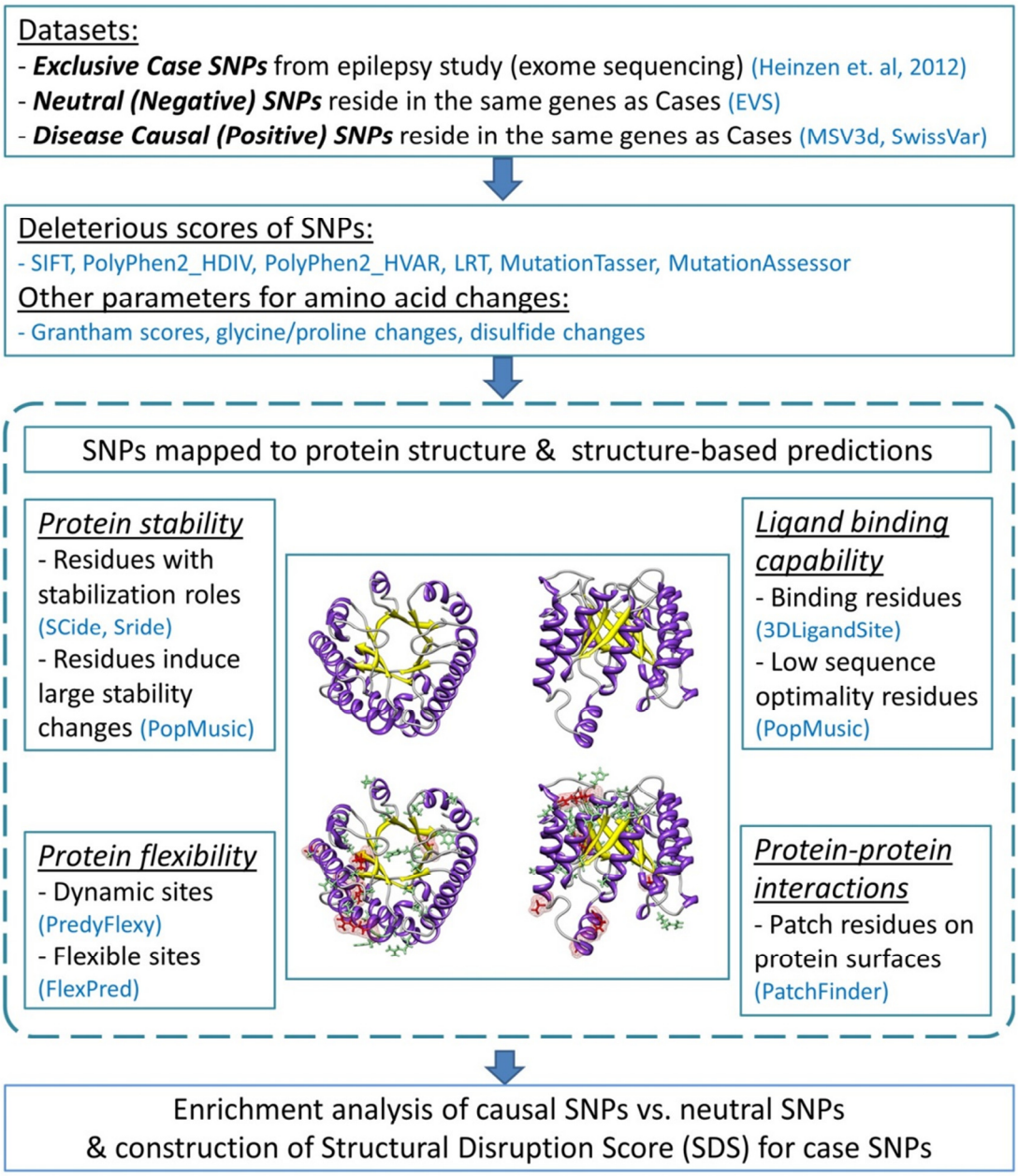


Figure 4.1: Flow diagram of the analysis pipeline. The analysis employed sequence-based deleterious prediction scores, parameters which reflect the nature of amino acid changes, and 3D structure-based evaluations. Structural analyses were performed by characterizing functionality of mutated protein residues caused by negative and positive SNPs (indicated by green and red stick representations, respectively). All analysis results were collectively used to evaluate enriched features found predominantly in causal SNPs. We then examined these predictors with regard to the case variants. The number of deleterious structure predictions per substitution represents a "structural disruption score" (SDS), and was used to rank candidate epilepsy variants.

Genomic dataset and candidate protein sequences

The epilepsy-specific amino acid substitutions identified from a recent exome sequencing study of epilepsy disorders [111] served as our case variants for which we aimed to assess whether or not they are likely to impact protein function. In that study, exome sequencing was performed on 118 cases and 242 controls. Follow-up genotyping for candidate causal variants included approximately 90% and 65% of individuals with European ancestry in the case (n=878) and control (n=1,830) groups, respectively [111]. The study identified 72 homozygous variants (68 are nsSNPs) found in 71 genes ("gene set 1") that were exclusive to cases. Among these, 52 nsSNPs were present in more than one affected individual. All genes in this first dataset had been previously characterized but not known to cause epilepsy; therefore, we added a second gene set ("gene set 2") to represent genes known to associate with the disorders. We attained the second gene list (n=41 genes) from two public repositories of genetic variations: MSV3d [43] and SwissVar [44]; none of the genes overlap with entries from the primary dataset. There are 373 missense variants in the 41 genes that have been documented to cause epilepsy; therefore, we treated them as case variants for gene set 2.

For both sets of genes, we compiled corresponding negative neutral and positive causal variants from the EVS database (retrieved March 201) [50], and MSV3d (July 2012 release) [43] and SwissVar (accessed February 2013) [44], respectively. Positive controls are documented non-epileptic disease-causing nsSNPs found in the same genes (n=134 nsSNPs from 14 genes of set 1, and n=205 nsSNPs from 41 genes of set 2). Likewise, negative controls are variants observed in these genes, but with no clinical associations (neutral nsSNPs). Any negative controls already identified as either case or positive SNPs

were excluded from the list of neutral SNPs, resulting in 5,281 and 1,490 putatively neutral (i.e. negative control) SNPs for sets 1 and 2, respectively.

Gene and variant annotations

In order to infer amino acid indices for the altered amino acid residues, nsSNPs were mapped to their corresponding protein sequences and structures using transcript IDs. All protein sequences (major isoforms) were downloaded from the UniProt database (accessed February 2013) [51]. Prior to applying our new variant analysis protocol, we performed literature searches on the genes and SNPs in our datasets in order to manually annotate their influence on the disease. In particular, we compared the features of gene sets 1 and 2, and recorded relevant findings.

First we grouped genes by their related biological pathways or biological functions using a gene group profiling method [78]. Second, we performed literature searches using SNPshot—a text mining tool for PubMed abstracts (accessed December 2012) [80]. Third, we assumed that amino acid mutations caused by the rare case SNPs or the causal SNPs would locate in the vicinity of functional sites of protein chains. Therefore, we utilized UniProt's sequence feature records (accessed February 2013) [51] to check if the mutating amino acids locate in any of the important sites; e.g., molecule processing sites, binding sites, modification sites, etc.

Population-specific minor allele frequencies (MAFs) for all variants were compiled from NHLBI GO Exome Sequencing Project (ESP6500) (June 2012 release) [50], available from dbNSFP 2.0 (accessed March 2013) [15].

Protein structure dataset

We used protein 3D structures to determine the structural nature of altered protein residues and to evaluate the effects of single point mutations introduced by nsSNPs on a specific protein. To ensure that we represent most of the proteins with high quality 3D structures, we employed several structural sources. Experimentally derived structures were retrieved from RCSB (retrieved April 2013) [60]. Homology models were compiled from SAHG (retrieved July 2013) [113] or automatically built using Phyre2 (accessed April 2013) [101]. Multiple structural candidates representing an overlapping protein chain were compared and only one best structure was chosen to represent the best non-overlapping protein segment.

Details of the two approaches for acquiring protein homology models are as follows. First, we searched for 3D models from the SAHG database [113], which contains a collection of encoded human protein structures, constructed by Modeller software [114]. We downloaded only structures having >15% sequence identity to the template. The retrieved proteins exhibit either ligand bound (holo) and/or unbound (apo) forms. Second, we built protein models by multiple template methodology using the automated Phyre2 homology modeling server [101]. Structure templates were selected by default and models were built from variable numbers of high confidence templates. This multi-template approach ensures that the model covers most of the protein chain. Large proteins (>1,300 amino acids) were truncated into smaller domain(s) using domain boundary information from InterPro [115]. A model representing each shortened sequence was built independently using either the single- or multi-template method; there was no attempt to join multiple models into a single model for a protein. For models

created with the Phyre2 server, we retained the best homology model based on the empirical criteria that >50% of the residues were modeled at >90% confidence.

After the initial homology model selection, the models were further subjected to energy minimization with explicit solvent using the YASARA force field [62] to resolve any steric conflicts found within the structures. Next, we validated the homology models using two independent scores: QMEAN6 [63] and ModFOLD4 [64]. Both scores show good ability to distinguish between good and bad models in the recent Critical Assessment of protein Structure Prediction (CASP) experiments [116]. To facilitate the structural validation step, we selected structures that pass the QMEAN6 threshold for subsequent ModFOLD4 evaluations.

In many cases, we initially selected more than one validated structure to represent an identical protein domain. To retain only one best representative structure for a protein segment, we used Chimera [65] to visualize all structure candidates and determined the structural similarity among them using two parameters: root-mean-square deviation (RMSD) of C_α atoms, and quality score (Q-score) that normalizes an RMSD by the alignment length. All measurements were performed with Chimera's MatchMaker tool [65]. When several overlapping structures agreed with each other, we selected the one with the best ModFOLD4 score. When the structures were in disagreement, we discarded them all together. Our retrieval and validation pipeline for protein 3D structures yielded 114 non-overlapping structures representing 57 gene products from gene set 1, and 51 non-overlapping structures representing 36 proteins from gene set 2.

Table 4.1 summarizes the number of missense variants from our genomic dataset in three categories (case, negative, and positive controls), with respect to the presence/absence of their corresponding 3D structures.

Table 4.1: Number of variants within each gene set, classified into three classes (cases, negative controls, and positive controls), and numbers of 3D structures used in the analysis.

Gene set (# of genes)	Number of variants by categories*			Number of 3D structures by types and sources**					# of selected structures	# of genes with selected structures
				Crystal structures	Homology models					
	Case	Neg	Pos	RCSB	Phyre2 (multi-template)	Phyre2 (single-template)	SAHG (apo)	SAHG (holo)		
Set 1 (71)	30 (68)	1674 (5281)	100 (134)	20 (24)	8 (35)	35 (59)	20 (86)	31 (80)	114	57
Set 2 (42)	184 (373)	554 (1490)	105 (205)	2 (2)	3 (19)	5 (17)	21 (46)	20 (38)	51	36

* Number of variants by categories is indicated by the number of SNPs locate within the set of selected 3D structures (114 structures for gene set 1, and 51 structures for gene set 2), followed by the total number of SNPs with and without 3D structures (number shown in parentheses).

** Number of structures represents the number of selected 3D structures that passed quality validation scores. The initial number of structures obtained from each data source is much larger, indicated by numbers in parentheses.

Inferring variant deleteriousness from sequence-based predictors

We obtained sequence-based predictions for each amino acid variant from dbNSFP 2.0 (accessed March 2013) [15]. The program provides pre-computed deleteriousness scores for six established deleterious prediction algorithms: SIFT [16], PolyPhen2_HumDiv and PolyPhen2_HumVar [19] , LRT [17], MutationTaster [20], and MutationAssessor [18]. Three evolutionary conservation-based scores were also included: GERP++ [53], phyloP [54], and SiPhy [55]. For simplicity, we assigned the level of deleteriousness and

conservation to each mutation based on how many predictors reported the mutation to be either “deleterious” (maximum score of 6) or “conserved” (maximum score of 3).

Additional parameters for sequence-based analysis

In addition to the SNP-based prediction parameters that are derived from multiple sequence alignments, some useful information can be analyzed from a single protein sequence alone. For example, amino acids with similar physicochemical properties may substitute for one another while maintaining the functionality of the protein. Three indicators may be used to highlight the most severe changes of amino acid pairs. First, Grantham scores [56] reflect the degree of physicochemical difference between pairs of amino acids. Second, changes involving any glycine or proline residues are likely to affect protein function since these two residues have special roles with regard to protein structure: proline has an exceptional conformational rigidity compared to other amino acids while glycine is much more conformationally flexible [117]. Third, gain or loss of disulfide bonds occurs when variants induce changes in cysteine residues. Disulfide bond formation between non-adjacent cysteines can facilitate protein folding; hence, they are important for maintaining the structural integrity of the protein [118]. In this context, we used DiANNA webserver [119] to predict the disulfide connectivity patterns in the wild type protein, and then determined if the amino acid mutation affects the bonding of cysteine pairs.

Inferring variant deleteriousness from structure-based predictors

A number of currently available protein structural analysis tools have the potential to be applied to structure-based variant assessment protocols [34]. To assess the functionality

of mutated protein residues, we concentrated on four features of structural analysis: protein stability, protein flexibility, protein-ligand binding potential, and protein-protein interaction potential. Many mutations disrupt these elements, and as a result, contribute to disease etiology.

1) *Protein stability.* For assessment of protein stability, we aimed to first identify amino acids with specialized roles in promoting protein stability, and second to determine which mutations cause a significant change in protein stability. For the first objective, we used SCide webserver [69, 70] and SRide program [71] to identify amino acids with essential stability functions. Long-range stabilization center (SC) residues are pairs of amino acids having close atomic contact (sum of van der Waals radii <1 Å), but locate at least ten amino acids apart on the primary sequence [69, 70]. A subset of SC residues may make distinct contributions to protein stability because they are also evolutionary conserved and located in the core region of the protein, and/or have many interacting partners. SC residues with these two extra properties are referred as stabilizing residues (SRs); they are also expected to make key contributions to protein stability [71]. For the second objective, we aimed to determine if a particular mutation affects protein stability by means of inducing a large magnitude of free energy change ($\Delta\Delta G$). For this purpose, we selected PoPMusic 2.1 [120] as our $\Delta\Delta G$ predictor.

Amino acid changes that increase protein stability ($\Delta\Delta G < 0$) and those associated with the destabilizing mutation ($\Delta\Delta G > 0$) are noted. Due to large differences in performance of stability change calculations [121], the proper margin for severe stability change can be ambiguous. However, it is known that the sensitivity in predicting stabilizing

mutations is much less than for destabilizing ones [66], and the correlation between predicted stability change ($\Delta\Delta G_P$) and measured values ($\Delta\Delta G_M$) of our selected program is ~1 kcal mol^{-1} [120]. Therefore, in our study, we followed the suggestions made by Dehouck et al. (2011). The stability changes are categorized into four levels: no change if $\Delta\Delta G$ is between ±0.5 kcal mol^{-1}, mildly stabilizing if $\Delta\Delta G$ is between -0.5 and -2 kcal mol^{-1}, mildly destabilizing if $\Delta\Delta G$ is between 0.5 and 4 kcal mol^{-1}, and strongly destabilizing if $\Delta\Delta G$ is $\geq$4 kcal mol^{-1}.

2) *Protein flexibility.* Protein flexibility is an important protein feature because highly dynamic sites are often involved in special functions, such as binding residues that can undergo subtle motion rearrangements when a small molecule is bound. Flexible amino acid residues permit large protein movements during protein folding and conformational switches [122]. For evaluating the levels of residue dynamics within a protein, we employed the predicted B-factors (relative vibrational motion) and root-mean-square fluctuations (RMSFs) obtained from a prediction program PredyFlexy [75] to classify amino acid residues into rigid, intermediate or flexible sites. For predicting protein movements of higher amplitudes, such as in conformational switches, we used the program FlexPred [76, 77] to determine which amino acid residues are located at conformationally flexible sites, indicated by a probability value P(Flexible).

3) *Protein-ligand binding potential*. For a SNP that causes an amino acid change in the vicinity of a catalytic site or a ligand binding site, it is possible to determine whether the mutation is indeed affecting the catalytic activity or the ligand binding affinity of the protein. In silico predictions are possible, but they require extensive computational

resources. We utilized two alternative approaches to predict the ligand binding sites or catalytic sites from protein 3D structures, and assessed whether or not the altered protein residues locate in or near the predictions. The first approach began with the use of 3DLigandSite [72] to search for ligands present in homologous structures. Then, a cluster of amino acids located within a default distance setting of 0.8 Å of the selected ligand was predicted as a pocket site, and amino acid residues that make up that pocket site were specified as ligand binding residues. In the second method, catalytic sites were predicted by scanning for protein residues that are not well optimized. This assumption is based on the finding that catalytic sites are generally designed for function rather than stability [120]. Low optimality residues are those whose several possible mutations would improve protein stability. The program PoPMusic [120] is fast enough that it can calculate stability changes ($\Delta\Delta$Gs) of all possible mutations at a given position in the protein sequence, and was used to identify non-optimized amino acid residues based on the summation of all stabilizing $\Delta\Delta$Gs. This parameter designates the degree of non-optimality (Γ) for each amino acid residue along the protein chain.

4) *Protein-protein interaction potential.* Disease-causing mutations that do not occur in binding sites or buried sites are predominantly found on protein interfaces [36]; therefore, we assessed which of the mutating protein residues may be involved in this type of inter-molecular interaction. We used PatchFinder program [73, 74], which evaluates evolutionary conservation scores in conjunction with solvent accessibility of protein residues to determine the most significant cluster of conserved residues on the surface of a protein. This patch indicates possible functional sites of protein-protein interactions.

Additional parameters for structure-based analysis

Other information retrieved from the structure-based data includes the type of protein secondary structure that each variant interrupts, and the relative solvent accessibility of the altered protein residue. We obtained these predictions from PoPMusic [120]. Due to the small sample size, we modified the eight reported types of protein secondary structure [123] into five groups: (G/H/I), E, (T/B), S, and C. The 3-, 4- and 5-turn helices (groups G, H and I, respectively) were grouped jointly as helices. Extended strand in parallel and/or anti-parallel β-sheet remains as an individual group (group E). Groups T or B correspond to helices or sheets whose hydrogen bonding patterns are too short to form proper secondary structures. Lastly, groups S and C denote bend and coil annotations, respectively.

The relative solvent accessibility (RSA) for residue X is expressed as a percentage of that observed for an Alanine-X-Alanine tripeptide. This conformation would expose the central X residue in the tripeptide as much as would normally be possible in a protein [120]. We considered protein residues whose RSA ≤20% as buried sites. Otherwise, they are expected to be on the protein surface.

Statistical comparison of positive and negative SNPs

We tested which predictions and measures can statistically distinguish between negative and positive control SNPs from each gene group. Particularly, we assessed which characteristics are most likely to be enriched in positive controls, and therefore imply disruption of functionality. After defining thresholds of likely deleterious function, we classified the predicted values for negative and positive controls into each different

category. The categories for structural indicators, the numerical cutoff values, and the numbers of mutations with extreme values from the two datasets are summarized in **Table 4.2**.

Table 4.2: Categories for structural indicators, cutoff values for continuous numerical parameters, and number of SNPs with extreme measures. All cutoff values were defined exclusively for gene set 1. The counts and percentages of variants with extreme values in each of the three variant classes are included.

Indicators	Cut-offs	#Case (%) (n=30)	#Neg (%) (n=1674)	#Pos (%) (n=100)
Stability change	*Stabilizing*: $\Delta\Delta G$ between -2 to -0.5 kcal mol $^{-1}$	0	19 (1%)	0
	Strong stabilizing: $\Delta\Delta G \leq$ -2 kcal mol $^{-1}$	0	0	0
	Destabilizing: $\Delta\Delta G$ between 0.5 to 4 kcal mol $^{-1}$	20 (65%)	830 (50%)	66 (66%)
	Strong destabilizing: $\Delta\Delta G \geq$ 4 kcal mol $^{-1}$	0	1 (0%)	0
Dynamic sites	*Highly rigid*: B-factor$_{norm} \leq$ -0.537 (@2.5 percentile)	0	38 (2%)	3 (3%)
	Highly dynamics: B-factor$_{norm} \geq$ 1.17 (@97.5 percentile)	1 (3%)	44 (3%)	1 (1%)
	Highly rigid: RMSF$_{norm} \leq$ -0.607 (@0.5 percentile)	0	48 (3%)	2 (2%)
	Highly dynamics: RMSF$_{norm} \geq$ 1.195 (@99.5 percentile)	1 (3%)	48 (3%)	3 (3%)
Flexible sites	*Conformationally rigid*: P(Flexible) $\leq$ 0.158 (@2.5 percentile)	1 (3%)	34 (2%)	2 (2%)
	Conformationally flexible: P(Flexible) $\geq$ 0.860 (@97.5 percentile)	1 (3%)	52 (3%)	0
Sequence optimality	*Highly non-optimal*: $\Gamma \leq$ -5 kcal mol $^{-1}$	1 (3%)	49 (3%)	5 (5%)

For continuous parameters, we compared the difference between the means of the positive and negative controls using two-tailed unpaired t-tests (**Table 4.3**). The distributions of scores within each SNP group were also illustrated by density plots (**Figures 4.2-4.3**). For non-numerical characteristics, we used Fisher's exact test to determine whether the proportions of negative and positive control SNPs for each of the features are significantly different (**Table 4.4**). For continuous measures, similar analyses were performed after first transforming the scores into discrete categories based on pre-specified thresholds that are most likely to discriminate normal and aberrant residues.

Once a series of predictions and measures was generated for all possible variants in a gene set, we converted continuous parameters into categorical classes, utilizing both literature-based and empirical cutoff values to represent the extremes (**Table 2**).

Table 4.3: T-test statistics for gene sets 1 and 2. All tests were performed on a subset of SNPs whose protein structures pass quality validations. Significant statistics indicate different means between negative and positive SNPs.

Parameters		Prob > |t|, (t ratio), df†					
		Set 1 (57 genes)			Set 2 (36 genes)		
Sequence-based deleterious scores	SIFT	<.0001*	(6.60)	df 1704	<.0001*	(4.61)	df 598
	PolyPhen2_HDIV	<.0001*	(8.15)	df 1772	<.0001*	(7.38)	df 657
	PolyPhen2_HVAR	<.0001*	(10.22)	df 1772	<.0001*	(8.70)	df 657
	LRT	<.0001*	(4.19)	df 1734	.0066*	(2.73)	df 654
	MutationTaster	<.0001*	(9.39)	df 1674	<.0001*	(5.57)	df 631
	MutationAssesssor	<.0001*	(15.30)	df 1772	<.0001*	(6.06)	df 652
Sequence conservation scores	GERP	<.0001*	(5.30)	df 1772	<.0001*	(3.92)	df 657
	phyloP	<.0001*	(4.81)	df 1772	0.0010*	(3.29)	df 657
	SiPhy	<.0001*	(5.89)	df 1771	<.0001*	(4.95)	df 657
Structure-based scores	ΔΔG	<.0001*	(4.81)	df 1772	<.0001*	(3.83)	df 657
	B-factor	.0190*	(-2.35)	df 1756	.2450	(1.16)	df 657
	RMSF	.2304	(-1.20)	df 1756	.2264	(1.21)	df 657
	P(Flexible)	.1185	(-1.56)	df 1772	.7594	(-0.31)	df 657
	Γ	.9090	(-0.11)	df 1772	.2308	(-1.20)	df 657
	RSA	.0035*	(-2.93)	df 1772	.0658	(-1.84)	df 657

† Statistic parameters include the two-tailed p-value, value of the t-statistics (t ratio), and the degree of freedom (df). Significant p-values (α =.05) are designated with '*'. The number of "df" equals to *n-2* samples used in the analysis.

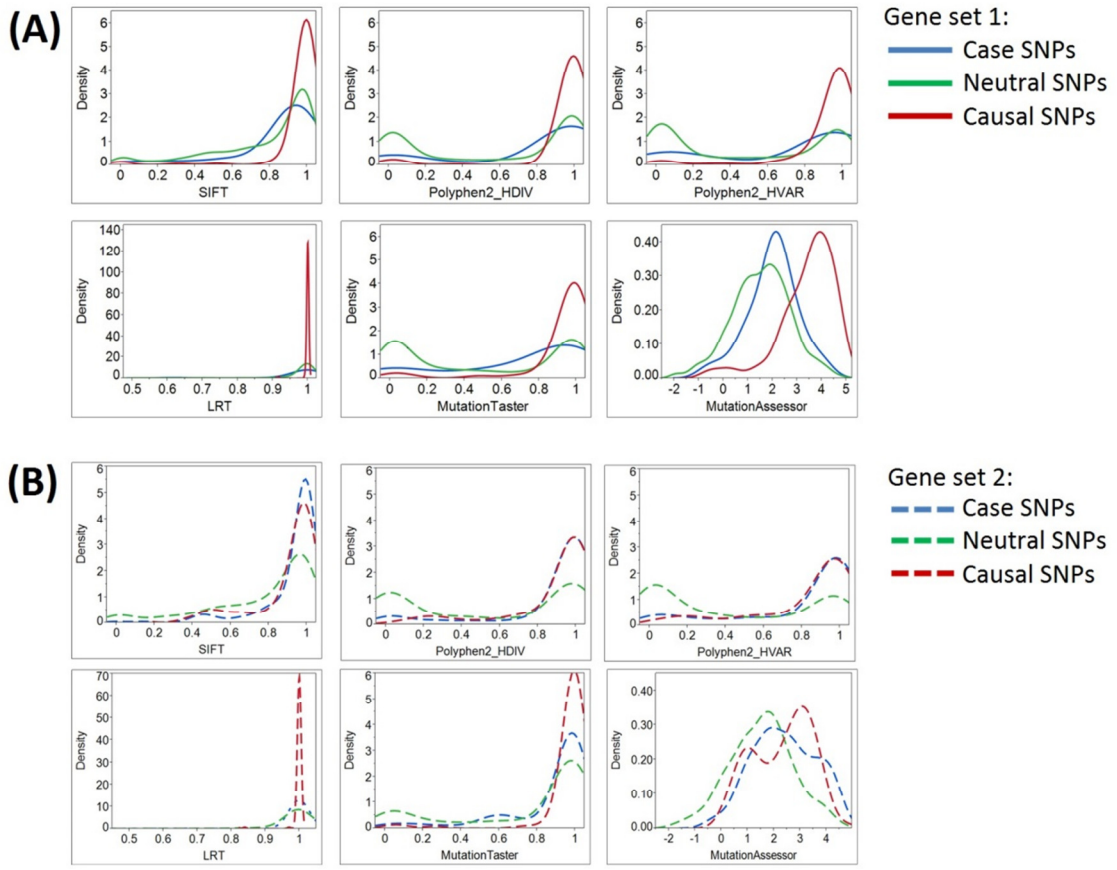


Figure 4.2: Density plots of six deleterious scores for Case, Neutral and Causal SNPs. By most of the standard deleteriousness scores, the distributions of cases in gene set 1 (Panel A) are closer to the neutral than the causal variants, and the neutral and causal variants are significantly different. The "known epilepsy" dataset (gene set 2, Panel B) demonstrated similar results. In this gene set, variants documented to cause epilepsy were regarded as "cases", while variants associated with other disease types were considered as "causal SNPs (positive control)". Although three prediction programs (SIFT, Polyphen2_HDIV, and Polyphen2_HVAR) suggested case and causal SNPs share similar distributions of deleterious scores, the remaining three programs illustrate their prediction algorithms do not favor the two types of causal variants equally. More importantly, case SNPs (epilepsy-causing SNPs) resemble neutral SNPs more than the causal ones.

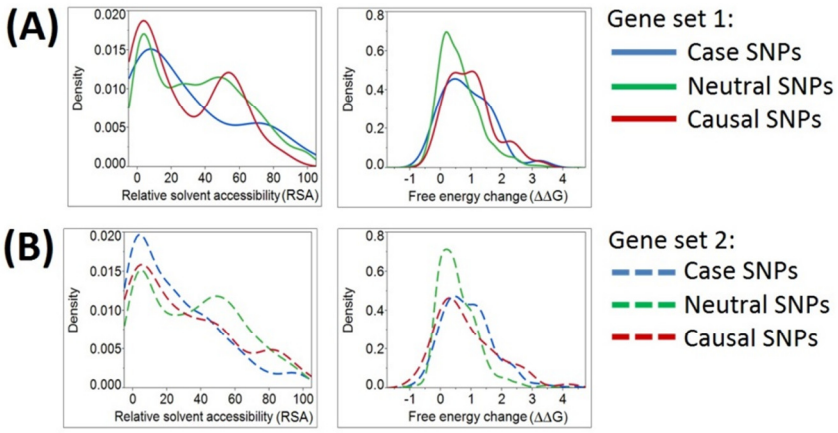


Figure 4.3: Density plots for relative solvent accessibility and free energy change for Case, Neutral and Causal SNPs. The two structure-based scores demonstrate that Case and Causal SNPs share similar characteristics. Results were obtained from two independent sets of genes (Panels A-B). Note the shift of the blue curves (cases) toward the causal SNPs (red) and away from the neutral ones (blue).

Table 4.4: Fisher's exact test statistics for gene sets 1 and 2. High sequence conservation (conservation count = 3) is a feature that enriched in causal SNPs of gene set 2, but is more likely to be found in neutral SNPs of gene set 1.

Enrichment types	Significant features	Fisher's exact test (one-tailed)†	
		Set 1 (57 genes)	Set 2 (36 genes)
Enriched in causal SNPs	Deleterious count ≥ 4	< .0001*	< .0001*
	Conservation count = 3	Enriched in neutral SNPs	< .0001*
	Induce large amino acid change (Grantham score ≥ 100)	< .0001*	< .0001*
	Induce disulfide change	NS (p = .6081)	< .0001*
	Induce gly/pro change	.0003*	< .0001*
	Locates in buried site (RSA ≤ 20%)	.0109*	< .0001*
	Locate in conformationally rigid site (P(Flexible) ≤ 0.74)	NS (p = .0558)	.0060*
	Locate on protein patch	< .0001*	< .0001*
	Locate in protein domain	.0204*	< .0001*
	Strongly reduce protein stability ($\Delta\Delta G \geq 4$ kcal mol $^{-1}$)	NS (p = 1)	.0400*
	Reduce protein stability ($\Delta\Delta G \geq 0.5$ kcal mol $^{-1}$)	.0009*	< .0001*
Enriched in neutral SNPs	Conservation count = 3	< .0001*	Enriched in causal SNPs
	Locate in highly dynamics site (B-factor$_{norm}$ ≥ 97.5 percentile)	NS (p = .2671)	.0224*
	Locate in highly flexible site (P(Flexible) ≥ 97.5 percentiles)	.0468*	NS (p = .1432)

† Data for gene set 1 includes 100 causal SNPs and 1674 neutral SNPs. Data for gene set 2 includes 289 causal SNPs (184 epilepsy case variants plus 105 non-epilepsy positive control variants), and 554 neutral SNPs. Significant p-values (α =.05) are designated with '*'. Non-significant test statistics are labeled with NS, followed by the correspondent p-value.

For protein stability and sequence optimality measures, we followed the suggested thresholds from the PoPMusic program [120]. Stability changes were classified into four groups (mildly stabilizing, strongly stabilizing, mildly destabilizing, and strongly destabilizing) using the aforementioned cutoffs for ΔΔG. Regarding sequence optimality, it has been shown that as the threshold for Γ becomes more stringent, the proportion of catalytic sites to other sites increases [120]. For our analysis, we selected a cutoff for

which residues having $\Gamma \leq -5$ kcal mol $^{-1}$ are more likely to locate in ligand-binding domains.

For the remaining continuous measures, i.e., B-factor, RMSF, and P(Flexibility), we used empirical criteria to define the extremes. Extreme values for data set 1 and 2 were derived independently, but with an identical approach. First, we compared the score distributions of each measurement, collected from all possible SNPs in the protein set. Then, we selected the thresholds for each parameter so that we captured a handful of extreme variants. When applicable, we made sure that our thresholds do not induce large numbers of misclassifications. In summary, our empirically-defined thresholds are generally set at the top and bottom 2.5 percentiles. B-factor and RMSF predictions were classified into either highly rigid (extreme small negative values) or highly dynamic (extreme large positive values). Similarly, we denoted residues as conformationally rigid if the P(Flexibility) is exceptionally low, or conformationally flexible if the probability is notably high.

Assigning a structural disruption score to candidate epilepsy variants

After testing for statistically significant differences between negative and positive SNPs, we summarized the list of deleterious structure predictions and examined these predictors with regard to the case variants. We counted the number of deleterious structure predictions per substitution and represented this number as a "structural disruption score" (SDS). The scores were ranked and candidate epilepsy variants with a score of ≥ 4 out of 7 are suggested to be "putative structural disrupted variants". Further partitioning of this list based on the gene's tolerance of polymorphism (RVIS) [124] yields two subgroups:

variants of high tolerance genes (genes that have more variants than expected), and variants of low tolerance genes (genes that have less variants than expected). These two groups of variants are also discussed in detail with respect to their disease implications.

To examine the contribution of each selected parameter, especially the sequence-based deleterious score, towards our SDS, we compared the values of SDS with the Condel composite score [22], derived from three of the deleteriousness measures (SIFT, PolyPhen, and MutationAssessor). The evaluation was performed with a step-wise procedure. First, we tested for a correlation between the Condel score and SDS—including all parameters from the sequence-based and structure-based predictions (total n parameters). Then, we re-evaluated the correlation using *n-1* parameters, by excluding one of the SDS components at a time.

Results

Candidate gene and variant annotations

Despite the fact that Heinzen et al. (2012) performed pathway analysis on 1183 genes harboring either homo- and/or heterozygous nsSNPs, with a genotype exclusive to the case group, no significantly over-represented biological pathways were found [111]. Using an alternative gene group profiling method [78], we also did not observe a statistical abundance for any biological terms derived from gene set 1. However, this method did reveal that ~40% of genes in set 2 have roles in transmission of nerve impulses, ion channel complexes, or ion gated channel activity. A text mining method [80] discovered only 1 gene from set 1 that may be linked to epilepsy. This gene codes for a ubiquitin-like modifier activating enzyme 2 (UBA2), a drug metabolizing enzyme

that plays roles in GABAergic and cholinergic neuronal development [125]. Specifically, mutations in ubiquitin protein ligase along with disruptions in the important neuronal GABA receptor genes are suggested to induce seizure [126]. By contrast, all genes in set 2 are suspected to be involved in a wide range of epilepsy disorders [43, 44]. Also note that the proportion of variants in cases relative to controls is much lower for genes in set 1 than in set 2 (**Table 4.1**).

Annotation of altered amino acid residues indicated some consistent patterns between case and causal variants. When we performed sequence feature searches [51] to compute the number of variants localizing in structurally or functionally important sites of a protein chain, we observed that more than half of the positive SNPs in both gene sets 1 and 2 were predicted to locate in transmembrane, topological domain, or repeat regions. Similar patterns were found for the case-exclusive epilepsy variants.

Statistically significant differences between positive and negative SNPs

Table 4.3 documents that all deleteriousness scores, all conservation scores, and some of structure-based scores have significantly different means between negative and positive SNPs. The single parameters that best differentiate the groups are the MutationAssesssor prediction for gene set 1 and the PolyPhen2_HVAR prediction for gene set 2, but more notable is the highly significant differentiation for all of the sequence-based scores. Although the t-ratios are consistently lower in set 2 than set 1, the smaller sample of genes precludes inference that there is a difference between the two sets. Notably, three of the structure-based measures are also at least nominally significantly different between positive and negative control variants in set 1, and trend in the same direction in set 2:

ΔΔG protein stability, B-factor protein flexibility, and relative solvent accessibility (RSA).

We converted several of the continuous structural measures to categorical 'normal' versus 'extreme' values and compared profiles of disease causal variants with those of neutral variations in gene sets 1 and 2. **Table 4.4** reports the list of significantly distinct sequence/structural features for each variant category based on Fisher's exact test. Seven significant characteristics of causal variants found in the 57 genes in set 1 are: having a high deleterious count (≥ 4 out of 6 scores), introducing an amino acid change with large physicochemical dissimilarity, inducing glycine or proline change, being situated in a protein domain, buried site or on a conserved protein surface, and causing at least mild protein destabilization. By contrast, negative control SNPs of this gene set were found to be enriched in conserved variants (conservation count = 3 out of 3) and generally locate in conformationally flexible sites.

These findings were validated with a parallel analysis of gene set 2, although in this case we gained statistical power by combining the set of epilepsy case variants with the non-epilepsy positive control SNPs (combined disease-SNPs n=289). Each of the significant features detected in gene set 1 replicates in set 2, and additional evidence that disruption of disulfide bonds and location in conformationally rigid sites differentiate neutral and disease variants were obtained. These findings emphasize that causal variants are likely to affect protein functional sites, a conclusion that can only be obtained from structure-based analysis.

Structural features predict deleteriousness of case SNPs

Next, we asked how the distribution of scores for the putative epilepsy-case variants compares with the negative and positive controls. By most of the standard deleteriousness scores, the distributions of cases are closer to the negative than the positive controls in both datasets (**Figure 4.2**). We conclude that there is little evidence from standard measures for enriched deleteriousness in the case variants from the epilepsy study. Similar observations were found for Grantham score, protein flexibility parameters, and a few other structural measures (data not shown).

However, we determined that the solvent accessibility measure (RSA) and a protein stability measure ($\Delta\Delta G$) assign case variants to be more comparable to positive than negative controls (**Figure 4.3A**). Likewise, the same two structural parameters place "known epilepsy" variants in gene set 2 (**Figure 4.3B**) closer to the distribution observed in other disease-causing mutations. This analysis emphasizes the potential for structure-based deleteriousness measures to generate predictions that are more discriminating than those derived from measures of sequence conservation.

To obtain a list of high-priority functional nsSNPs for epilepsy, we applied the deleterious structure predictions enriched in positive SNPs (**Table 4.4**) to all candidate epilepsy variants and identified the ones with high SDSs (score ≥ 4 out of 7). A list of 14 high-likelihood structure-disrupted variants from 30 missense mutations was generated. To account for differences in the burden of mutations among genes, we used the Residual Variation Intolerance Score (RVIS) [124] to identify and compare the levels of mutational intolerance of each gene in our two datasets. The parameter determines the

deviation of observed vs. expected numbers of common variants in a gene. Petrovski et al. (2013) found that genes which carry many common mutations (large positive RVISs) are less likely to influence disease development. Comparison of average RVIS between genes in sets 1 and 2 indicated that the two groups do not have the same tolerance to variations (p-value .0011, two-tailed t-test). The average RVIS for gene set 1 is 0.39 (range 0.11 to 0.67, 95% CI) whereas the value for gene set 2 is -0.40 (range -0.77 to -0.03, 95% CI). As expected, lower RVISs were observed for the documented disease causal genes (gene set 2). The finding is consistent with low RVISs in many OMIM genes [124]. Nonetheless, the positive average RVIS for gene set 1 is not surprising; among the 57 genes in set 1, 38 genes (67%) are classified as being high tolerance.

Sub-classification of our 14 high SDS structure-disrupted variants yields 9 and 5 genes that are highly acceptable or tolerant of mutations, respectively (**Table 4.5**). Although variants in genes with low tolerance (negative RVISs) are more likely to be deleterious, our structural disruption score focuses at the variant level, and the structural analysis potentially provides novel intuition that is not apparent from any sequence- or gene-based analysis. Therefore, the indication of many "high tolerance genes" in our dataset does not preclude potential functional effects of specific variants. For this reason, we performed in-depth literature searches on all 9 and 5 variants of the 2 subgroups, and provide our inference of the likelihood that a particular SNP may contribute to the epilepsy disorders (**Table 4.6**). Interestingly, we found that half (4 out of 9) of the SNPs in high tolerance genes have potential to promote epilepsy. The proportion is comparable (2 out of 5) for variants located in the low tolerance genes. Of the 9 variants in high tolerance genes, one has a structural feature that is compatible with those of neutral SNPs, i.e., the variant

alters a highly flexible protein site; therefore, we disregarded it as a putative functional variant. Similar consideration of the low tolerance genes suggests that just one, *PPP1R27*, is likely to harbor a mutation that promotes epilepsy. This leads to prioritization of 9 "high-priority putative functional variants for epilepsy". The locations of each of these variants with respect to their protein 3D structures are shown in **Figure 4.4**, and each is discussed below (variant numbering follows **Table 4.5**).

Table 4.5: Case SNPs with high structural disruption scores. Each candidate epilepsy variant is suggested to disrupt protein structure/function if its structural disruptions score (SDS) is high (≥ 4 out of 7). A list of 14 putative structural disrupted variants from 30 missense mutations is reported, along with the corresponding 7 scores that make up the SDS. The variants are classified into two groups, based on the level of polymorphism tolerance of its gene.

Variant category	List	Gene	Variant position (base change, [amino acid change])	Structural disruption features*							SDS (max = 7)	Additional gene/SNP features	
				High deleterious count? [Del count, max=6]	Large amino acid change? [Grantham score]	Induce Gly/Pro change? [amino acid change]	Locate in buried site? [%RSA]	Locate on protein patch?	Locate in protein domain? [CATH architecture]	Reduce protein stability? [ΔΔG (kcal mol^{-1})]		Gene tolerance level [RVIS]	Locate in highly flexible site [percentile of P(flexible)]
structural disrupted variants which locate in high tolerance genes	A	*ABCA6*	(4075)TGC>CGC [C1359R]	yes [6]	yes [180]		yes [2.05]		yes [3-layer(αβα) sandwich]	yes [1.64]	5	high [0.26]	
	B	*ABHD14A*	(685)CGA>GGA [R227G]		yes [125]	yes [R>G]			yes [3-layer(αβα) sandwich]	yes [0.87]	4	high [0.77]	
	C	*ALOX12*	(1211)CGG>CAG [R404Q]	yes [4]			yes [10.55]	yes	yes [up-down bundle]	yes [0.90]	5	high [0.80]	
	D	*DDX52*	(1064)ATC>ACC [I463T]	yes [4]			yes [4.42]		yes [3-layer(αβα) sandwich]	yes [1.37]	4	high [0.05]	
	E	*EPYC*	(449)TCC>TGC [S150C]	yes [5]	yes [112]		yes [0.77]		yes [α-β horseshoe]		4	high [0.51]	
	F	*HELB*	(1517)GAT>GGT [D506G]	yes [4]		yes [D>G]	yes [18.27]			yes [1.40]	4	high [1.08]	
	G	*IAH1*	(127)CTG>GTG [L43V]	yes [4]			yes [4.66]		yes [3-layer(αβα) sandwich]	yes [2.06]	4	high [0.17]	
	H	*NMUR1*	(409)CGC>TGC [R137C]	yes [4]	yes [180]		yes [1.17]		yes [up-down bundle]	yes [1.83]	5	high [0.27]	
	I	*PALB2*	(2993)GGA>GAA [G998E]	yes [5]		yes [G>E]	yes [2.27]			yes [3.23]	4	high [0.32]	yes [97.75]

Table 4.5 (continued)

Variant category	List	Gene	Variant position (base change, [amino acid change])	Structural disruption features*							SDS (max = 7)	Additional gene/SNP features	
				High deleterious count? [Del count, max=6]	Large amino acid change? [Grantham score]	Induce Gly/Pro change? [amino acid change]	Locate in buried site? [%RSA]	Locate on protein patch?	Locate in protein domain? [CATH architecture]	Reduce protein stability? [ΔΔG (kcal mol^{-1})]		Gene tolerance level [RVIS]	Locate in highly flexible site [percentile of P(flexible)]
structural disrupted variants which locate in low tolerance genes	J	*EXOG*	(830)GGA>GTA [G277V]	yes [6]	yes [109]	yes [G>V]			yes [3-layer(αβα) sandwich]	yes [1.64]	5	low [-0.45]	
	K	*FAAH2*	(821)CGT>CAT [R274H]	yes [4]			yes [0]		yes [α-β complex]	yes [0.95]	4	low [-0.29]	
	L	*MAOA*	(374)AAT>AGT [N125S]	yes [4]			yes [6.65]		yes [orthogonal bundle]	yes [1.36]	4	low [-0.14]	
	M	*PPP1R27*	(336)ATA>ATG [I112M]	yes [5]			yes [0]		yes [α horseshoe]	yes [1.41]	4	low [-0.32]	
	N	*PTPN14*	(566)GAA>GGA [E189G]	yes [4]		yes [E>G]			yes [up-down bundle]	yes [1.79]	4	low [-0.30]	

* Only the values corresponding to enriched characteristics of causal SNPs are included in the table; the empty cells do have values but they are not presented here for clarity.

Table 4.6: Summary of structural disrupted case SNPs. The table summarizes the structural and clinical findings for each of the top 14 case variants. The 9 putative functional variants for epilepsy (variants A-H, M) are identified from of a subset of the 14 variants. Eight variants (variants A-H) are located in "high tolerance genes" and do not possess a compatible feature with those of neutral SNPs. The ninth variant (variant M) is located in a "low tolerance gene"; the gene is likely to be associated with epilepsy disorders.

Variant category	List	Gene	Variant position (base change, [amino acid change])	Gene functions [biological function]	Variant's features*	Disease implications†		SDS (max = 7)	%MAF (AfrAmr, EurAmr)
						Other diseases	Epilepsy		
structural disrupted variants which locate in high tolerance gene	A	*ABCA6*	(4075)TGC>CGC [C1359R]	**ABC transporter A family member 6 [lipid homeostasis]**	large change in amino acid properties, mutation causes protein destabilization but does not alter disulfide bonds	n/a	n/a	5	0.30, 2.00
	B	*ABHD14A*	(685)CGA>GGA [R227G]	hydrolase [neuron development]	near active site (low Γ), loss of side chain (R>G)	linked to Chanarin-Dorfman syndrome (fat depositions in internal organs)	less likely	4	0.16, 0.23
	C	*ALOX12*	(1211)CGG>CAG [R404Q]	lipoxygenase [lipid metabolism]	near active site, on best protein patch	shares substrate with COX (COX-2 expression increases upon electrical stimuli)	likely	5	0.05, 0.37
	D	*DDX52*	(1064)ATC>ACC [I463T]	RNA helicase [mRNA degradation control]	stabilization center	gain of function in Drosophila's homolog suppresses seizure; mRNA loss accounts for 1/3 of human diseases	maybe	4	0.23, 1.16
	E	*EPYC*	(449)TCC>TGC [S150C]	epiphycan [cartilage development]	stabilization center, mutation yields preferred hydrophobic core	osteoarthritis	no	4	0.61, 2.42
	F	*HELB*	(1517)GAT>GGT [D506G]	DNA helicase [DNA damage repair]	3 indications as a binding residue, confirmed by mutagenesis	effective cellular protection mechanism helps animals survive brain injuries after induced seizures	likely	4	0.57, 3.76
	G	*IAH1*	(127)CTG>GTG [L43V]	esterase [lipid metabolism]	mutation locates at a turn region (not favorable in highly structured proteins)	antiepileptic drugs interfere with lipid metabolisms	likely (drug response)	4	0.28, 2.62
	H	*NMUR1*	(409)CGC>TGC [R137C]	neuromedin-U receptor 1 [uterus contraction, vasoconstriction]	diminishes stabilizing salt bridge and causes protein destabilization	control of food intake	no	5	0, 0.08
	I	*PALB2*	(2993)GGA>GAA [G998E]	partner and localizer of BRCA2 [homologous recombination repair]	mutation may interfere with conformational flexibility of protein, largely decreases protein stability	several mutations identified in breast cancer but disease associations are not definitive	no	4	0.59, 2.40

Table 4.6 (continued)

Variant category	List	Gene	Variant position (base change, [amino acid change])	Gene functions [biological function]	Variant's features*	Disease implications†		SDS (max = 7)	%MAF (AfrAmr, EurAmr)
						Other diseases	Epilepsy		
structural disrupted variants which locate in low tolerance gene	J	*EXOG*	(830)GGA>GTA [G277V]	mitochondria endonuclease [programmed cell death]	rigid residue at turn region, controls positioning of C-terminal and active site, confirmed by mutagenesis	n/a but reduces substrate binding	n/a	5	0.27, 1.20
	K	*FAAH2*	(821)CGT>CAT [R274H]	fatty-acid amide hydrolase2 [lipid metabolism]	mildly decreases protein stability	gene's regulation of endocannabinoid system is linked to Alzheimer's and other CNS disorders	maybe	4	n/a
	L	*MAOA*	(374)AAT>AGT [N125S]	monoamine oxidase type A [neurotransmitter metabolism]	far from functional sites, mildly reduces protein stability	gene catalyzes several neurotransmitters and associated with behavioral phenotypes, confirmed by animal studies	no	4	n/a
	M	*PPP1R27*	(336)ATA>ATG [I112M]	phosphatase regulator [cellular process regulation]	longer amino acid side chain causes steric clash	member of KEGG epilepsy pathway; protein in the same family linked to Lafora disease (teenager-onset of epilepsy)	likely	4	n/a
	N	*PTPN14*	(566)GAA>GGA [E189G]	tyrosine-protein phosphatase non-receptor type 14 [cellular process regulation]	mutation does not alter inter-residue bonding but slightly decreases protein stability	several mutations identified in colorectal cancers	no	4	0.93, 3.22

*Variant's features include all structural changes/implications that were collected during the analysis, regardless of their significant in feature enrichment towards causal SNPs.

†Disease implications denote any clinically-relevant associations in found in literatures. Epilepsy implications indicate our opinions on whether or not the variant contributes to epilepsy development. The opinion is based upon several data sources. However, the considerations exclude the SDS of a variant and its minor allele frequencies (MAFs). (The SDS had already been utilized as a filter during the variant prioritization step. The allele frequencies are presented here solely for comparison purposes.)

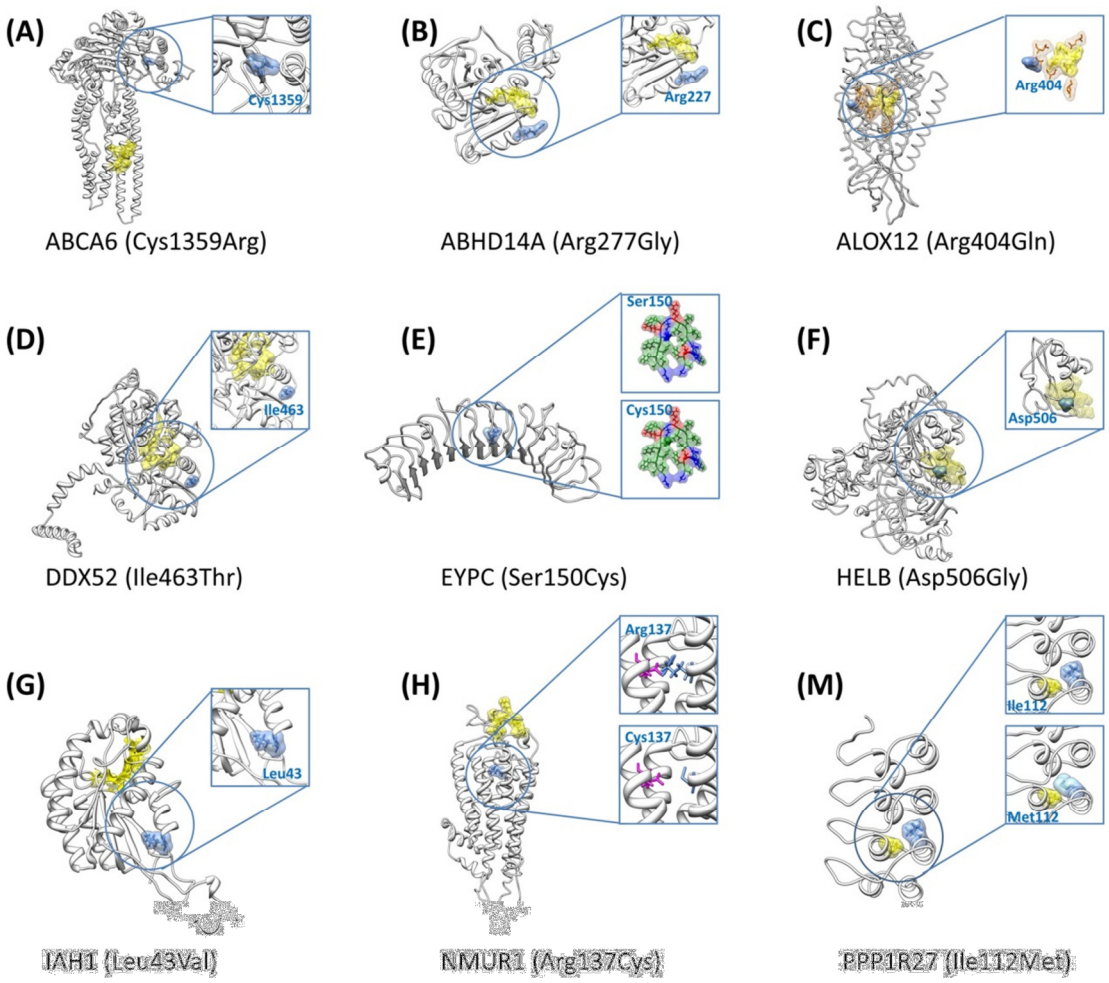


Figure 4.4: 3D structures for the nine high-priority variants for epilepsy. Mutated protein residues caused by the SNPs are indicated with blue surface representations along with the amino acid name (wild type) and residue index. Yellow surface representations refer to predicted binding sites or catalytic sites, except in **(M)** in which it represents a protein residue with close proximity to the altered amino acid. Orange representations in **(C)** indicate the predicted best conserved residue cluster on the protein surface. The substitution of *Ser150Cys* in **(E)** adds one favorable hydrophobic residue (green surface representation) to the core of the α/β horseshoe. The magenta sticks in **(H)** represent the partner residue, *Glu117*, which forms a salt bridge with wild type *Arg137*. This salt bridge is lost in the presence of mutant Cys at position 137.

Individual assessment of putative structurally deleterious variants

A) *Cys1359Arg in ABCA6* Cys1359 mutation is located in a buried site of ABCA6, a protein that plays roles in macrophage lipid homeostasis (**Figure 4.4A**). Despite the high SDS of this variant (score 5/7), it has not been associated with diseases. The functionality of cysteine residue depends largely on the protein structure and its cellular location. For this instance of Cys to Arg change, it does not alter the pattern of disulfide bonds, partly due to the rarity of this bonding type within membrane proteins [99]. However, the C1359R mutation is still considered as a crucial change (Grantham score 180, ΔΔG 1.64 kcal mol^{-1}), especially when the mutation occurs in the middle of protein domain.

B) *Arg277Gly in **ABHD14A.*** Arg277 is located on an exposed site of **ABHD14A** (**Figure 4.4B**). **Another** member of this protein family, **ABHD5,** is thought to be responsible for a rare genetic disorder called Chanarin-Dorfman syndrome. This is the only variant among the 14 structural disrupted case SNPs that was not detected as "deleterious" by any sequence-based algorithms (deleterious count 0/6), but our SDS suggests it has potential impact on the protein (SDS 4/7). **The wild type** Arg227 residue has low sequence optimality (Γ -0.58 kcal mol^{-1}), a likely indicator that it has close proximity to a catalytic site. Moreover, the Arg to Gly change is considered as an unfavorable substitution, especially when it occurs in a structured protein, as the lack of a side chain in Gly may diminish proper protein folding or intermolecular interactions.

C) *Arg404Gln in ALOX12* Arg404 is found in a buried site of ALOX12 (**Figure 4.4C**). Arg404 is in close proximity with the catalytic site of this protein and is also predicted to be part of the protein patch for intermolecular interactions. The Arg404Gln substitution within this protein is also predicted to cause a slight protein destabilization (ΔΔG 0.90

kcal mol^{-1}), despite the acceptable Grantham substitution score (< 100). We suspect the variant may play roles in epilepsy implication, since a link between its substrate (arachidonic acid) and seizure susceptibility had been proposed [127].

D) *Ile463Thr in **DDX52*** Ile463 is located in a buried site of DDX52 (**Figure 4.4D**). **A** study of seizure susceptibility in *Drosophila* discovered that a gain-of-function mutation in the *maleless* helicase gene can suppress seizure susceptibility in bang-sensitive *Drosophila* mutants [128]. **The** Ile463Thr mutation in this protein affects the structural integrity of the protein since it is detected as a stabilization center (SC), consistent with its destabilization effect (ΔΔG 1.37 kcal mol^{-1}).

E) *Ser150Cys in EYPC* The Ser150Cys variant is located in the core of the leucine-rich repeat (LRR) structural motifs of EYPC (**Figure 4.4E**). Several disulfide bonds are formed between cysteine clusters that flank the LRRs, providing additional structural support, but no changes of disulfide bonds or hydrogen bonds appear in the mutated protein. Ser150Cys has minimal impact on protein stability (ΔΔG = -0.41 kcal mol^{-1}), and while the substitution induces a large physicochemical change (Grantham score ≥ 100), the substitution is considered neutral if located in α/β protein [129]. Indeed, we observed Ser150Cys added one favorable hydrophobic residue to the core of the α/β horseshoe. Despite its high SDS (4/7), we consider it unlikely that this variant contributes to any disorders.

F) *Asp506Gly in HELB* Asp506 is situated in a buried site of HELB (**Figure 4.4F**) and several of its features are predicted to interfere with ligand binding. First, wild type Asp506 has an exceptionally low sequence optimality value (Γ -4.14 kcal mol^{-1}). Second, wild type Asp506 is predicted to be a highly flexible site by three parameters, although

the values are not extreme. Third, the Asp506Gly substitution is predicted to destabilize the protein (ΔΔG 1.40 kcal mol^{-1}). These characteristics coincide with a recent mutagenesis experiment that proves Asp506 is part of a binding motif, and the mutation of D506A induces loss of substrate binding when associated with E499A and D510A [130]. Study of another human DNA helicase, Twinkle, demonstrated that two missense mutations were detected in patients with a wide range of psychiatric symptoms, including severe epileptic encephalopathy, possibly due to inefficient recovery from molecular stress [131]. HELB Asp506 is thus particularly interesting for further assessment of a role in epilepsy.

G) *Leu43Val in IAH1* Leu43 is located at a buried site in IAH1, although it lies far from the substrate binding site (**Figure 4.4G**). The protein is of interest given that many antiepileptic drugs are potent enzyme inducers and inhibitors of the cytochrome P450 system, which affects lipid and glucose metabolisms, as well as evidence that increased lipase level is one of the side effects of anti-psychotic and anti-epileptic drugs [132, 133]. Substitution of Leu to Val is quite favorable in all protein folding types [129], but Leu43Val in this protein is suspected to reduce protein stability (ΔΔG 2.06 kcal mol^{-1}). A plausible explanation may be that this protein is highly structured, comprising only a few loop residues. The turn regions may play an essential role in bringing together and enabling or allowing interactions between regular secondary structure elements.

H) *Arg137Cys in NMUR1* Arg137 is situated in the central cavity of NMUR1 (**Figure 4.4H**). The presence of the a mutant Cys at position 137 diminishes the stabilizing salt bridge between wild type Arg137 and Glu117, and results in a decrease in protein stability (ΔΔG 1.83 kcal mol^{-1}). The gene has no known associations with any diseases.

M) *Ile112Met in PPP1R27* Ile112 is found at a buried site of PPP1R27 (**Figure 4.4M**). A study of a similar protein, PPP1R3C, identified one missense mutation that may **lead to a mild phenotype in Lafora disease**—a teenage onset epilepsy disorder [134]. In addition, protein phosphatase 1 (PP1) was identified as a member of long term potentiation (PTP) pathway in epileptogenesis and epilepsy (KEGG: map04720) [135, 136]. Other genes in this pathway mostly regulate neurotransmission and ion channel receptors. The Ile to Met substitution is quite favorable both in general (Grantham score < 100) and in distinct types of protein folding [129]. However, the longer aliphatic side chain of Met creates steric clashes with Ala104 of an adjacent helical region. This single point mutation is predicted to destabilize the protein ($\Delta\Delta G$ 1.41 kcal mol^{-1}).

Structural disruption score correlates with sequence-based deleterious score

Finally, we tested whether the SDS correlates with measures not used to construct the score itself. We performed an additional sequence-based deleterious prediction of case SNPs using Condel: a weighted score that integrates the output of five tools [22], three of which were used in our analysis. There is a significant positive correlation between our SDS and the Condel score (p-value .0342, n= 30). The trend persists after removing the sequence-based deleterious scores from SDSs of 16 variants (the variants have deleterious count ≥ 4 out of 6), although the significance in correlation is reduced to a marginal level (p-value .0717). In addition, we performed a similar analysis by sequentially removing one SDS component from the total score, and found the significantly positive correlation between SDS and Condel is maintained (**Table 4.7**). The weakest structure-based parameter among all of the SDS components is the classification of buried vs. exposed site using RSA. After removing this indicator from 15

variants, we did not detect any correlation between the two measures (p-value .13). Note that this analysis is complicated by the small sample sizes. Nonetheless, the finding supports our expectation that although the deleterious count is one of the major components for constructing our SDS, the remaining non-conventional deleteriousness parameters also have substantial impact on the overall missense variant evaluation.

Table 4.7: Step-wise analysis for correlation of SDS and Condel score. Initial SDS includes 7 parameters, described in **Table 4**; the maximum SDS for each variant is 7. The revised SDS calculates the score by exclude one SDS component at a time; the maximum revised score for a variant equals 6.

SDS parameters	R^2 of linear fit	P-value of correlation	ǂof variants affected by the revised SDS[†]
All SDS	.1504	.0342**	none
SDS-high deleterious count	.1112	.0717*	16
SDS-large amino acid change	.1508	.0340**	5
SDS-induce gly/pro change	.1308	.0496**	7
SDS-locate in buried site	.0782	NS (p = .1344)	15
SDS-locate on protein patch	.1629	.0270**	3
SDS-locate in protein domain	.2093	.0110**	25
SDS-reduce protein stability	.1243	.0560*	20

[†] The full dataset has 30 missense variants. All data points were used to test for a correlation between SDS and Condel score. When an SDS component was removed during the step-wise analysis, the SDSs for some numbers of variants were affected, i.e. the excluded parameter was applicable to the variant. For such cases, the correlation analysis was performed with all of the 30 data points, minus the number of exclusions indicated in the last column.

Significant p-values are designated with '**' and '*' for $\alpha = .05$ and $\alpha = .10$, respectively. Non-significant test statistics are labeled with NS, followed by the correspondent p-value.

Discussion

Current perspectives in prioritization of epilepsy variants

The evaluation and prioritization of candidate case variants is particularly difficult when the disorder involves a dissimilar set of genes, as is the case for epilepsy disorders, which are now known to involve diverse molecular pathways [112, 137]. The classic epilepsy genes (ion channel genes/regulatory genes, neurotransmitter genes/receptor gene/regulatory genes, genes that disrupt cortical circuits, and genes that lower the convulsion stimuli) rarely present in genomic data from sequencing studies. This may be because the classical epilepsy mutations are Mendelian, whereas exome sequencing likely targets more polygenic cases, noting that only 1% of epilepsy disorders are inherited in a Mendelian manner [138].

Ferraro and coworkers (2012) highlighted how daunting the task is for epilepsy when they exemplified some factors in the design of cohort studies that influences the discovery of true positive epileptic variants: the selection of cases (presence/absence of the cause of symptoms), the seizure types (determine the amount of genetic influences), the patient profiles (age at onset, gender, characters of seizure incidences, etc.), and the assumption of genetic patterns (common variant effects, rare variant effects, or a combination of both) [139]. Some authors are starting to incorporate disease information as prior knowledge in the probabilistic evaluation of candidate causal SNPs [105, 140, 141], but the scarcity of knowledge related to epilepsy genes limits this approach. Indeed, we found that the genes in our dataset are somewhat poorly understood and their disease contributions are largely unknown.

In addition to SNPs, structural variations have been shown to associate with epilepsy genetics. Jia et al (2011) used stepwise enrichment analysis of protein-protein interactions to derive a molecular network of 20 high priority candidate genes linked to copy number variation [142]. Interestingly, the genes do not overlay with the 68 homozygous nsSNP-containing genes in our dataset (but they do overlap with 4/1,604 genes that harbor heterozygous variants: [111]). An independent comparison with nine genes that harbor *de novo* mutations (identified from trios—unaffected parents and their affected child) also found no overlap [143].

Consequently, we cannot be sure that any of the variants discussed in this article are truly causal for epilepsy. However, the variant prioritization scheme does suggest a reduced number of candidates which, on the basis of careful curation of protein structure, might be taken forward for targeted experimental manipulation and assessment of biological function in cell lines or model organisms.

Key findings

We have developed a structure-based variant analysis protocol that evaluates the effects of missense mutations with respect to their predicted effects on protein features, such as solvent accessibility, stability, and flexibility. Replicated trends for putative case SNPs to have aberrant structural features that more closely match those of established disease mutations in the same proteins, than to those of neutral polymorphisms, establish the potential utility of this approach as an orthogonal protocol to sequence-based assessment of deleteriousness.

Starting with 71 genes harboring putative case SNPs from an exome sequencing study of epilepsy disorders [111], we were able to perform the assessments on 57 gene products. Presumably only a fraction of these are actually causal, so our expectation was simply that the distribution of risk scores may be shifted from neutral toward disease-associated. Several features were observed to classify SNPs into two groups: putative functional variants, or presumably neutral variants, and a composite risk score based on summation of these features highlighted nine putative functional variants, from thirty exclusive missense mutations whose protein 3D structures are available. Although none of these has been previously linked to epilepsy disorders, detailed case-by-case analysis strongly suggests that several should be prioritized for further functional evaluation.

Although our structure-based analysis only captured a fraction of variant residues due to the limited availability of 3D structures (Table 5.1), we show that 44%, 32%, and 75%, of case, negative, and positive variants from the first gene set are amenable to structure-based predictions. The equivalent percentages for set 2 genes are 49%, 37%, and 51%, respectively. More importantly, we were able to represent 84% of the proteins in our first set, and 86% in the second set, despite the difficulty in generating high quality structures. Our preliminary analyses suggest that it is not necessary to have structural data for all variants in order to construct the SDS. Since the sequence-based predictions of variants with structure data are similar to those obtained for all SNPs in the dataset (data not shown), we are confident that the conclusions from our structure-based variant assessment protocol can be extended to the complete SNP pools in each of the two gene sets.

With our combined sequence- and structure-based analysis pipeline, we discovered that some features are predominantly found among negative or positive SNPs. Structure-based parameters contribute as much as 50% of the features that differentiate the two types of variants. There are several key observations from our feature enrichment analysis. First, we noticed seven common characteristics that are predominantly found in the positive control SNPs, regardless of the set of genes. SNPs with strong effects are those that: have deleterious count ≥ 4, have Grantham score ≥ 100, induce glycine or proline changes, locate in protein domains, in buried sites, or on conserved protein patches, and destabilize the protein. Second, variants with neutral effects (negative SNPs) have a few strong enriched features. In gene set 1, we found negative SNPs are more likely to be in conformationally flexible sites. A similar feature was detected in gene set 2, in which non-damaging SNPs are mostly highly dynamic sites. An additional distinctive characteristic of the negative SNPs in gene set 1 is that they tend to affect highly conserved residues. This finding appears to be counter intuitive; we suspected that this unique observation seems to be an exception for this particular gene group.

Study limitations

A primary limitation of our approach is the requirement for homology models that support computational prediction of structural characteristics. Specifically, only 18/68 proteins had at least partial experimental structures, so homologous templates were used in most cases. These were not available for just 9% of the proteins (n = 6), but the retained models did not cover the disrupted site for another 56% (36/68 variant sites), and 24 of the potentially disrupted proteins are larger than 1000 amino acids, which also presents additional challenges for building models and satisfying quality settings.

Nevertheless, by restricting the modeling to domains, we were able to model 84% of the 68 candidate genes (covering 44%, 32%, and 75%, of case, negative, and positive variants). This is a clear improvement on the automated pipelines used in training algorithms such as PolyPhen2. We also ensured that the quality of the models was validated wherever possible, which also introduces an intensive manual curation requirement into the analytical pipeline, requiring some knowledge with methods that most genomicists are not familiar with.

An analytical limitation is that the size of the dataset is relatively small, since only one variant per gene was studied, and just 68 proteins were available to begin with. Since these are structurally diverse, it is likely that different aspects of protein structure are affected and the probability of enrichment for any one structural feature is correspondingly reduced. While approximately three quarters of amino acid changes leading to Mendelian diseases consistently induce protein destabilization [37], the structural consequences of missense variants in complex diseases such as the epilepsy disorders are likely to be of a more diverse nature.

Study innovations

The Structural Disruption Score (SDS) offers a novel strategy for genomic profiling of variants that have uncertain but likely weakly deleterious function. It combines similar instances of variants with respect to their predicted impact on aspects of protein structure, allowing joint assessment of the impact of the variants as a class on biological function. Instead of evaluating each variant one by one, SDS provides a ranking that might be used to guide downstream experimental and/or clinical evaluations.

Other studies using structural biology approaches to examine the variant effects have considered a large number of variants per gene, facilitating direct contrasts of variant characteristics and predictions for individual genes [144] or genes with similar structures/functions [39, 145]. The genomic data that we started with is relatively small by comparison. However, we provide an alternative structure-based approach that can accommodate the small number of variants expected from exome sequencing samples as well as the large diversity of gene functions they will typically generate.

Most importantly, our SDS score implementation assures that case variants share similar characteristics as those observed in causal variants, but not neutral variants known to reside in the same proteins. We have validated the findings in a replication dataset, and show how the approach can be used as a unique solution to prioritizing case variants in unrelated genes. Though not providing a guarantee of disrupted function, the score should be considered as a complementary approach to existing sequence-based deleteriousness prediction.

Conclusion

Using the list of enriched features, we concluded that this novel structure-based assessment protocol for missense variant deleteriousness has a potential to determine high-priority candidate variants suitable for experimental validations. The analysis may prove to be useful, particularly when traditional sequence-based predictions are inconclusive. An important question is whether the same structural attributes differentiate neutral and functional variants for different categories of disease.

Because our study employed large numbers of external resources (variant predictions, gene information, 3D structure modeling and quality controls, sequence- and structure-based predictions), the analysis pipeline presented here is not readily automated. Aspects of it are in theory readily generalizable to all classes of proteins, and once all the above steps have been accomplished, the variant deleteriousness structure-based predictions could be effectively populated into a database. After that labor-intensive step is completed, the SDS for any variants in a dataset may be computed and retrieved virtually by combining the predictions for the genes specific for the study.

Acknowledgments

This work was supported by start-up funds from the Georgia Tech Research Foundation to GG, and TP was supported by the School of Biology at Georgia Tech.

CHAPTER 5: SYSTEMATIC 3D SCREENING OF AMINO ACID MUTATIONS IN PHARMACOGENES

Abstract

Next-generation sequencing technologies have promoted steady progress in the identification of genetic factors that influence drug responses. Priority is being given to genes whose variants are predicted to alter pharmacokinetic and/or pharmacodynamic parameters, which may result in an increased risk of drug toxicity or therapeutic failure. Therefore, rational analysis of an individual's genomic variants can guide personalized medical treatments. In this study, we selected 48 genes identified as "Very Important Pharmacogenes (VIPs)" by the PharmGKB database. Despite their high potential impact for prescription of specific drugs, only ten gene-drug pairs currently have dosing guidelines suggested by the Clinical Pharmacogenetics Implementation Consortium (CPIC).

To promote more practical usage of genetic test results, we have developed a systematic screening for structural disturbance of amino acid mutations within the 48 VIPs in the context of their 3-dimensional (3D) protein structures. Our pipeline focuses on the changes in inter-residue bonding, protein stability, protein flexibility, drug binding capability, protein-protein interactions, and amino acid dissimilarity, in addition to the localization of the variants and the amino acid secondary structure preference. These results are incorporated into the construction of a five-feature "SDS Pharmacogenes" score that categorizes distinguishable characteristic profiles that annotate VIP variants as

functional rather than neutral mutations. Unlike most existing conservation-based measures, SDS Pharmacogenes can be used to evaluate unknown variants of these 48 VIPs and predict the degree to which each one will have strong impacts towards pharmacogenomics, potentially aiding optimization of drug therapy.

While expertise in 3D-protein analysis is beneficial, our implementation does not require that an individual with experience in protein structures be engaged in the personalized genome evaluation. In addition, the analysis pipeline is systematic and scalable, thus expected to keep pace with the rapid accumulation of pharmacogenomic data.

Introduction

Pharmacogenomics aims to predict the effect of genetic polymorphisms and mutations on therapeutic efficacy and/or toxicity. To this end, one of many challenges for bioinformaticians is the development of variant evaluation methods that may be used to guide the interpretation of genomic variation.

In this study, we surveyed genomic variability of 48 Very Important Pharmacogenes (VIPs) identified in the PharmGKB database [146] (**Table 5.1),** and implemented a novel 3-dimensional (3D)-protein structure based method to systematically screen the characteristics of their amino acid mutations. The implementation incorporates results from many protein structural analysis tools. It generates an integrative score that offers a comprehensive understanding of potential mutational effects. The analysis also reveals some key structural features that are generally present in functional variants (mutations that lead to functional differences between wild type and mutant proteins) as well as highlights the complexity of understanding mutation-function relationships in genes that

are highly interactive. Our findings may assist the refinement of existing deleterious prediction algorithms by adding protein structural data to sequence conservation based-parameters, particularly to predict the functional significance of variants in pharmacogenes. The objective is to provide a comprehensive understanding of potential mutational effects towards inter-patient drug responses and lead to the practical usage of genetic test results.

Table 5.1: List of 48 VIPs and number of their pharmacogenomics associated variants. The data were obtained from PharmGKB's curated list of clinical annotations. There are no variants in level 2B or level 4.

Gene	Protein name	CPIC	Total number of variants (Nonsynonymous/Synonymous/non-coding)				
			All levels	Level 1A	Level 1B	Level 2A	Level 3
ABCB1	Multidrug resistance protein 1	no	78 (53/7/18)			3 (3/0/0)	75 (50/7/18)
ACE	Angiotensin-converting enzyme	no	14 (0/2/12)			1 (0/0/1)	13 (0/2/11)
ADH1A	Alcohol dehydrogenase 1A	no					
ADH1B	Alcohol dehydrogenase 1B	no					
ADH1C	Alcohol dehydrogenase 1C	no					
ADRB1	Beta-1 adrenergic receptor	no	5 (5/0/0)				5 (5/0/0)
ADRB2	Beta-2 adrenergic receptor	no	8 (8/0/0)			1 (1/0/0)	7 (7/0/0)
AHR	Aryl hydrocarbon receptor	no					
ALDH1A1	Retinal dehydrogenase 1	no	1 (0/0/1)				1 (0/0/1)
ALOX5	Arachidonate 5-lipoxygenase	no	1 (0/0/1)				1 (0/0/1)
BRCA1	Breast cancer type 1 susceptibility protein	no					
COMT	Catechol O-methyltransferase	no	10 (6/0/4)			1 (1/0/0)	9 (5/0/4)
CYP1A2	Cytochrome P450 1A2	no	14 (0/1/13)				14 (0/1/13)
CYP2A6	Cytochrome P450 2A6	no	2 (0/0/2)				2 (0/0/2)
CYP2B6	Cytochrome P450 2B6	no	28 (27/0/1)			8 (7/0/1)	20 (20/0/0)
CYP2C19	Cytochrome P450 2C19	yes	107 (56/7/44)	16 (10/2/4)		46 (25/2/19)	45 (21/3/21)
CYP2C8	Cytochrome P450 2C8	no	8 (5/0/3)				8 (5/0/3)
CYP2C9	Cytochrome P450 2C9	yes	47 (39/0/8)	5 (5/0/0)	3 (3/0/0)	8 (8/0/0)	31 (23/0/8)
CYP2D6	Cytochrome P450 2D6	yes	213 (167/0/46)	34 (23/0/11)	8 (7/0/1)	98 (78/0/20)	73 (59/0/14)
CYP2E1	Cytochrome P450 2E1	no	4 (1/0/3)				4 (1/0/3)
CYP2J2	Cytochrome P450 2J2	no					
CYP3A4	Cytochrome P450 3A4	no	23 (8/0/15)			4 (1/0/3)	19 (7/0/12)
CYP3A5	Cytochrome P450 3A5	yes	24 (2/0/22)		1 (0/0/1)	2 (0/0/2)	21 (2/0/19)
DPYD	Dihydropyrimidine dehydrogenase	yes	14 (10/0/4)	3 (2/0/1)		1 (1/0/0)	10 (7/0/3)
DRD2	D(2) dopamine receptor	no	10 (1/1/8)			1 (0/0/1)	9 (1/1/7)

Table 5.1 (continued)

Gene	Protein name	CPIC	Total number of variants (Nonsynonymous/Synonymous/non-coding)				
			All levels	Level 1A	Level 1B	Level 2A	Level 3
F5	Coagulation factor V	yes	1 (0/0/1)			1 (0/0/1)	
G6PD	Glucose-6-phosphate 1-dehydrogenase	no	7 (3/0/4)		1 (1/0/0)	3 (1/0/2)	3 (1/0/2)
GSTP1	Glutathione S-transferase P	no	9 (9/0/0)			3 (3/0/0)	6 (6/0/0)
GSTT1	Glutathione S-transferase theta-1	no	5 (0/0/5)				5 (0/0/5)
HMGCR	3-hydroxy-3-methylglutaryl-coenzyme A reductase	no	9 (0/0/9)			1 (0/0/1)	8 (0/0/8)
KCNH2	Potassium voltage-gated channel subfamily H member 2	no	1 (0/1/0)				1 (0/1/0)
KCNJ11	ATP-sensitive inward rectifier potassium channel 11	no	1 (0/0/1)				1 (0/0/1)
MTHFR	Methylenetetrahydrofolate reductase	no	27 (27/0/0)		1 (1/0/0)	4 (4/0/0)	22 (22/0/0)
NQO1	NAD(P)H dehydrogenase	no	2 (2/0/0)				2 (2/0/0)
NR1I2	Nuclear receptor subfamily 1 group I member 2	no	1 (0/0/1)				1 (0/0/1)
P2RY1	P2Y purinoceptor 1	no	1 (0/1/0)				1 (0/1/0)
P2RY12	P2Y purinoceptor 12	no	6 (1/1/4)				6 (1/1/4)
PTGIS	Prostacyclin synthase	no					
PTGS2	Prostaglandin G/H synthase 2	no	3 (0/0/3)				3 (0/0/3)
SCN5A	Sodium channel protein type 5 subunit alpha	no					
SLC19A1	Folate transporter 1	no	6 (5/1/0)				6 (5/1/0)
SLCO1B1	Solute carrier organic anion transporter family member 1B1	yes	35 (21/0/14)	1 (1/0/0)		8 (4/0/4)	26 (16/0/10)
SULT1A1	Sulfotransferase 1A1	no	1 (1/0/0)				1 (1/0/0)
TPMT	Thiopurine S-methyltransferase	yes	10 (5/0/5)	4 (3/0/1)			6 (2/0/4)
TYMS	Thymidylate synthase	no	8 (0/0/8)				8 (0/0/8)
UGT1A1	UDP-glucuronosyltransferase 1-1	yes	13 (5/0/8)			4 (2/0/2)	9 (3/0/6)
VDR	Vitamin D3 receptor	no	3 (2/0/1)				3 (2/0/1)

Table 5.1 (continued)

Gene	Protein name	CPIC	Total number of variants (Nonsynonymous/Synonymous/non-coding)				
			All levels	**Level 1A**	**Level 1B**	**Level 2A**	**Level 3**
VKORC1	Vitamin K epoxide reductase complex subunit 1	yes	16 (2/0/14)	1 (0/0/1)	3 (0/0/3)	6 (1/0/5)	6 (1/0/5)
GSTP1	Glutathione S-transferase P	no	9 (9/0/0)			3 (3/0/0)	6 (6/0/0)
Total	48 genes	yes =10	776 (471/22/283)	64 (44/2/18)	17 (12/0/5)	204 (140/2/62)	491 (275/18/198)

The 48 VIPs have important roles in drug metabolism. Some mutations are expected to influence drug efficacy and/or toxicity through perturbation of the drug-binding ability of the protein (directly or indirectly), or via altered drug metabolism and excretion [147]. The PharmGKB database [146] (accessed January 19, 2014) lists 546 drug molecules that interact with one or many of the VIPs (**Supplementary Table B.1**). However, little is known about the function of most of the genetic variants. Despite their high potential impact for prescription of specific drugs, only ten gene-drug pairs currently have genotype-specific dosing guidelines developed by the Clinical Pharmacogenetics Implementation Consortium (CPIC) [148] (**Table 5.1**). These shortcomings highlight the need for a novel strategy to examine and predict the outcomes of genetic variations towards a diverse range of drug toxicity or therapeutic failure.

Several research groups have focused on expanding variant assessment methodologies for proteins, generally by including more predictive features [13, 35-39]. One of the notable improvements arises from adding predictive parameters which are derived from protein structural context. Even so, the space of protein study is large and many unexplored fields are worthy of attention [34]. Our study explores a broad range of protein analyses and employs statistical tests to derive key structural features of functional variants. The implementation utilizes numerous computational tools in the following eight areas of protein analysis: inter-residue bonding, protein stability, protein flexibility, drug binding capability, protein-protein interactions, residue localization, amino acid dissimilarity, and amino acid secondary structure preference. A subset of 11 tests which represent disrupted protein characters that are preferably targeted or induced by functional variants was identified among the total of 68 tested features. The initial selection of 11 significantly

enriched features is subsequently reduced to the 5 strongest indicators that can discriminate pharmacogenomics variants from neutral mutations.

The implementation of our fully structure-based variant assessment algorithm enhances our knowledge of residue features that are susceptible to or induced by functional mutations. Structure-based information provides both the understanding of local changes, as well as those that occur non-locally and cannot be deduced from sequence conservation analysis alone.

Methods

This research aims to provide a rational analysis of genetic variants in 48 Very Important Pharmacogenes (VIPs). The implementation of our 3D screening approach was designed such that it is systematic and scalable so that it offers a broad understanding of various protein characteristics while being flexible enough to accommodate emerging genomic data for future refinement.

Dataset of 48 Very Important Pharmacogenes (VIPs)

A list of VIPs was populated from the PharmGKB's external data source on October 7, 2013. In addition, we retrieved a list of curated gene-drug and gene-chemical pairs for each VIP from the gene's record, available at the PharmGKB website (accessed January 19, 2014) [146]. **Supplementary Table B.1** provides the descriptions of the 48 VIPs, their molecular functions, and the numbers of drug partners.

Genomic variation dataset

A) *Genomic variants for construction of the algorithm (training set)*

Genomic variants were obtained from publicly available databases and were categorized into two groups: functional mutations and neutral mutations. The data sources for attaining functional mutations include MSV3d [43] (accessed October 25, 2013), SwissVar [44] (accessed October 25, 2013), ClinVar [149] (accessed January 9, 2014), and UniProt's functional mutagenesis records [150] (accessed March 4, 2014). For neutral variants, we obtained the data from EVS [151] (accessed January 4, 2014) and Uniprot [150] (accessed January 28, 2014). Quality control of the genomic data was carried out by aligning wild type amino acids from the given list of variants to the corresponding UniProt curated protein sequences [150]. Any variants of minor protein isoforms were excluded (n=87).

Using the sequence mapping, we identified nine incorrectly labeled variants (one functional and eight neutral mutations). The variants have interchanges between reference and alternate residues. For example, a functional variant *ADRB2:c.79C>G (p.Gln27Glu)* is listed as a risk factor for obesity, metabolic syndrome and asthma in the ClinVar database (classified by a single submitter) [149]. The amino acid variant has a sequence conflict with our data sources: Uniprot [150] and SwissVar [44] list Glu as the wild type amino acid at residue 27, but ClinVar (based on dbSNP [49] indicates the SNP induces a change from Gln to Glu. These errors are likely due to inaccuracies of sequencing or in the reference genomes; therefore, all unmatched variants were eliminated from our study.

B) *Genomic variants for testing the utility of the algorithm (test set)*

A collection of pharmacogenomic-associated variants was downloaded from the "Clinical Annotation" records at the PharmGKB website [146] (accessed February 16, 2014). PharmGKB annotates the variant-drug associations based on the number and/or strength of pharmacogenomic evidence (**Table 5.1**). The two conditions classify variants into four levels: 1A/1B, 2A/2B, 3, and 4. Level 1A/1B variants are the most significant variants for pharmacogenomics. They have either endorsed CPIC guidelines or clinical implementations (level 1A variants) or strongly suggestive evidence of variant-drug associations from multiple cohort studies with significant test statistics (level 1B variants). Level 2B variants are those with moderate indications of variant-drug associations, but may lack statistical significance of associations in some studies. For any 2B variants that are located in VIPs, their association levels are re-assigned to 2A. Genomic variants in level 3 require more replication to confirm the significant test statistics from a single study, or the data from multiple studies is available but has not provided clear indication of an association. Other variants that were derived from case-by-case reports, non-significant studies, or *in vitro* data are assigned to be in level 4.

There are a total of 96 associations in level 1A/1B, 81 are found in 10/48 VIPs (*CYP2C19, CYP2C9, CYP2D6, CYP3A5, DPYD, G6PD, MTHFR, SLCO1B1, TPMT,* and *VKORC1*). Non-VIP genes that harbor strong pharmacogenomic associations are *CFTR* and *HLA-B* for a total of six associations in level 1A, and *ANKK1, CYP4F2, EGFR, GRIK4, IFNL3, TMEM43,* and *XPC* for a sum of nine associations in level 1B. The level 2B associations are related to variants of 55 non-VIP genes.

Among the 48 VIPs, there are 776 pairs of pharmacogenomic associations (**Table 5.1**).The associations are listed in the format of variant-gene-drug partners. Some gene-drug pairs are well recognized, while other genes may not directly relate to the drug's mechanism of action. To account for the accuracy of the annotations, especially for level 3 variants, we checked if their genes are listed as pharmacogenomic biomarkers in the US FDA drug labeling (http://www.fda.gov/drugs) (accessed March 7, 2014), namely, drug exposure and clinical response variability, risk for adverse events, genotype-specific dosing, mechanisms of drug action, and polymorphic drug target and disposition genes. Additional annotations for drug labels (informative/actionable pharmacogenomic labels from FDA and European Medicines Agency (EMA)) were obtained from the PharmGKB's website [146] (accessed March 7, 2014). In addition to the apparent pharmacogenomic evidence, some genes may have direct/indirect effects on drug efficacy or toxicity. The gene may be a drug target, enzyme, transporter, or carrier. We obtained this information from the DrugBank database [152] (accessed March 7, 2014). The clarifications of gene-drug pairs returned 307/776 associations with at least one of the listing drugs having FDA- or EMA- approved labels, (maximum 7 drugs/variant; CYP2D6 gene) and 548/776 listings in DrugBank [152] indicating that the gene has a known pharmacological action.

Of all 776 associations, 471 are related to missense variants (**Table 5.1**); their pharmacogenomic associations can also be in different levels due to multiple drug partners. The absolute number of unique missense mutations among the 471 associations, regardless of the association levels, is equal to 81. (For example, *Arg150His* of CYP2C19 has 10, 25, and 21 associations at levels 1A, 2A and 3, respectively (**Table 5.1**) but the

unique count of this mutation is one (one and the same variant). Since 9 of the 81 unique variants do not lie in regions of overlap with the protein structure dataset, we were able to use 386/471 structural data points to assess the abundance of enriched disrupted protein features with respect to the PharmGKB's three levels of pharmacogenomics association confidence. The equivalent 386 data points were used to evaluate the classification performance of six standard conservation-based predictors.

Protein structure dataset

We retrieved experimentally derived structures from the RCSB Protein Data Bank [60] and generated/downloaded numerous homology models following our previously described protocols [110]. We used two parameters, QMEAN6 [63] and ModFOLD4 [153], to validate all protein models.

Most (36/48) of the proteins in our dataset are well studied, granting access to multiple experimental structures (range 1 to 53 structures per protein). Nonetheless, almost all structures only partially represent the entire protein length. Among these PDB coordinates, we selected the best structural representative for each protein chain by comparing their resolution and the percentage of sequence coverage. To ensure that our 3D structural analysis covers most of the protein chain; we compared each of the selected PDB structures with the matching best homology models (from single-, or multi-template methodology, preferably full length structures). The best overall structure for each non-overlapping protein segment, among all sources, was retained for subsequent 3D analyses.

Note that some of the selected x-ray crystal structures contain sequence conflicts (from mutagenesis during crystallization) with respect to the reference protein sequences. For such cases, the structures were mutated back to the wild type proteins. We used the Swapaa tool in Chimera [65] to model the new side chain using the most probable amino acid rotamer [154]. Unfavorable atomic contacts between each modeled side chain with the remaining protein residues were assessed using the Find Clashes/Contacts tool in Chimera [65] with the default distance settings.

A complete list of protein 3D structures used in this study, along with the structure sources and quality parameters are provided in **Supplementary Table B.2.** The genomic and structure datasets were combined into 3D structure maps for each genetic variant. **Table 5.2** summarizes the genomic variability data of the 48 VIPs and the structure coverage statistics of the 3D maps. The complete list for genomic variability coverage of each gene and the percentage coverage of variants with structural data is presented in **Supplementary Table B.3.**

Table 5.2: Statistics of genomic variability data of the 48 VIPs and 3D structure maps.

Variant types	All genomic data*		Genomic data with available 3D structures		
	# of variants	Avg # of variants per gene	# of variants	Avg # of variants per gene	% of variant retained per gene†
Functional mutations	779 (in 35 genes)	16 (range 0-186)	371 (in 31 genes)	8 (range 0-58)	80%
Neutral mutations	2537 (in 48 genes)	54 (range 4-245)	1811 (in 45 genes)	38 (range 0-140)	84%

*Sources for functional mutations: MSV3d [43], SwissVar [44], ClinVar [149] databases and Uniprot databases [150]. Sources for neutral mutations: EVS [151] and Uniprot databases [150].

†Represents the average of all 3D variant coverage percentages of the 31 genes and 45 that harbor functional and neutral mutations, respectively.

Identification of conserved protein domain families

Protein domain families are evolutionary conserved structural modules which exist independently from the rest of the protein chain. Large proteins often consist of several structural domains. Protein domains were identified according to the annotations provided by InterPro—a centralize database for protein family (Hunter 2011). We downloaded the "All UniProtKB proteins" dataset (January 27, 2014 release, accessed February 16, 2014) and extracted domain definitions for the 48 proteins. Identification of protein domain family was performed by first indexing the mutated protein residues relative to the full length protein chain, and then identifying the protein domain family that surround each variant residue.

The InterPro database is made up from 11 external data sources [155], each with different identification and annotation methodologies for protein domains. We excluded the Gene3D-derived entries from the preliminary InterPro hits since Gene3D domains reflect the geometric similarity of protein structures rather than the functional or evolutionary relationships employed by the remaining algorithms. The exclusion of Gene3D hits yielded 4,599 remaining records; many variants have duplicate domain annotations from several algorithms. We further removed all duplicates to generate a non-redundant set of 113 protein domain families present in the 48 proteins (**Supplementary Table B.4**).

To detect protein domain families that harbor significantly more proportions of functional variants, we compared the relative abundance of functional and neutral variants that are located in a particular protein domain family. The calculations were performed by first normalizing the absolute number of functional (neutral) variants observed in each protein

domain family to the total number of functional (neutral) variants located within any domains (excluding variants found in domain boundaries). Each normalized value was then transformed into a percentage with respect to the sum of the two normalized values. The final two percentages were compared, and three types of domain family were identified: domains with more abundance of functional variants (n=36), domains with more abundance of neutral variants (n=42), and domains with only neutral variants (n=35).

Characteristics of amino acid mutations at different protein regions

1) Changes in physicochemical properties of amino acids

We performed Fisher's exact tests to examine the ratios of dramatic amino acid changes when the variants affect important vs. general protein regions. First, we used Grantham scores [56], which measure chemical dissimilarity between amino acid pairs, to classify amino acid changes into two classes. An amino acid replacement is considered conservative (mutation with similar amino acid property) if the Grantham score is less than 100 and radical (mutation with distinct amino acid property) otherwise. We used Grantham score to compare the amino acid changes at non-structural sites vs. structural sites [150], protein domains vs. domain boundaries [155], and binding sites (or around binding site) vs. other sites [72]. Second, we used the categorical classes of the amino acid side chain properties [156] to represent the four physicochemical changes between amino acid pairs: hydropathy index (hydrophobic and hydrophilic properties), volume, charge, and hydrogen donor/acceptor. The changes are empirically considered important if they occur between hydrophobic (A, C, I, L, M, F, W, V) and hydrophilic (R, N, D, Q,

E, K) side chains, between very small (A G, S) and large (R, I, L, K, M) or very large (F, W, Y) side chains, between very large (F, W, Y) and small (N, D, C, P, T) or very small (A, G, S) side chains, between positive charged (R, H, K) and negative charged (D, E) side chains, or between hydrogen donor (R, K, W) and acceptor (D, E) side chains. We tested the abundance of each altered side chain property at various protein locations: buried sites, exposed sites, binding sites, around binding sites, and any sites.

The Uniprot's sequence feature records provide descriptions of 11 functional features within a protein sequence. The annotations were either experimentally derived or computationally predicted. The features are: molecular processing, amino acid modification, sequence motif, sequence bias, active site, bond, natural variant, mutagenesis (indicates the amino acid has experimental mutation data), transmembrane, protein domain, and other region of interest. We re-organized the annotations into two groups: structural site and non-structural site. Our intuition is that the first four features do not necessarily have direct implications for the structural integrity of the protein; any variant annotation falling within these terms should be regarded as "non-structural site". Our genomic data were not annotated to have the fourth feature, sequence bias. The content of the first three most prominent features detected in our variant dataset are propeptide and signal peptide; motif, zinc finger, and repeat; and modified residue, respectively.

The grouping of missense variants into protein domain vs. domain boundaries was performed according to the aforementioned steps for identifications of protein domain families. Similarly, the partitioning of variants into either predicted binding sites or other sites was carried out at the subsequent step of prediction of ligand binding sites. We first

searched for ligand environments of each protein using 3DLigandsite [72]. The binding pocket residues were predicted from clusters of bound ligands that are present in protein homologs (average predicted binding site = 21 residues per protein, range 17-25 residues, 95% CI). There are 41 functional and 104 neutral variants predicted to lie within binding pockets; these variants were grouped into "predicted binding sites". Any other variants were assigned to the "other sites" category. Using the C_α- C_α atomic distances, we extracted a list of variant residues that are within 5 Å or 10 Å to the nearest predicted binding site [72]. The selected variants were grouped into either the "near binding sites (5 Å)" or the "near binding sites (10 Å)" categories, and their structural features were compared with all other variant residues.

2.) Changes in secondary structure preference of amino acids

Amino acids have different preferences towards a certain protein secondary structure (coil, strand, 3-turn helix, α-helix, bend, or turn). Malkov et. al (2008) compared the occurrence of an amino acid in all types of secondary structures and defined the amino acid as a preferred or a prohibited residue for that structure type [157]. We use this information in conjunction with the predicted secondary structure of a variant residue [120] to identify if the altered amino acid will induce a secondary structure break (i.e., the amino acid changes from preferred to inhibited for the predicted secondary structure).

Systematic 3D screening for structure-function relationships of amino acid mutations

We implemented a systematic 3D screening methodology to investigate structure-function relationships of each amino acid mutation. We concentrated the analyses on essential characteristics for a protein to function properly (independently or when

interacting with other molecules). The screening explores eight aspects of protein function: inter-residue bonding, protein stability, protein flexibility, drug binding capability, protein-protein interactions, residue localization, amino acid dissimilarity, and amino acid secondary structure preference. A similar set of computational tools was adopted in our previous study of epilepsy variants [110]. The compendium of scores was used to predict structurally-related attributes for each amino acid residue in the context of its 3D structure (**Table 5.3**). The following sections explain several assessments that have been added to this 3D screening process for the first time. The new methods include predictions of salt bridges and long range electrostatic interactions and hinge point predictions.

Table 5.3: List of computational tools used to analyze structurally-related attributes for each amino acid residue.

Structural feature	Parameter	Variable type	Description	Tool
Inter-residue bonding	Disulfide bond	Categorical	The mutation occurs at one of an amino acid pair that forms a disulfide bond.	DiANNA [119]
	Salt bridge / electrostatic interaction	Categorical	The mutation occurs at one of an amino acid pair that forms a salt bridge or an electrostatic interaction.	WHAT IF [158]
	Stabilization center (SC)	Categorical	The mutation occurs at one of an amino acid pair that forms an SC (two residues locate at least ten amino acids apart on the primary sequence but form close atomic contacts).	SCide [69, 70]
Protein stability	Stabilizing residue (SR)	Categorical	The mutation occurs at one of an amino acid pair that forms an SR (a subset of SC that is also evolutionary conserved and located in the core region of the protein, and/or have many interacting partners).	SRide [71]
	$\Delta\Delta G$	Categorical (transformed from continuous value)	Stability change from a single point mutation; categorized into four groups: no change if $\Delta\Delta G$ is between ±0.5 kcal/mol, mildly stabilizing if $\Delta\Delta G$ is between -0.5 and -2 kcal/mol, mildly destabilizing if $\Delta\Delta G$ is between 0.5 and 4 kcal/mol, and strongly destabilizing if $\Delta\Delta G$ is ≥4 kcal/mol.	PoPMusic 2.1 [120]
Protein flexibility (small amplitude)	B-factor$_{Norm}$	Categorical (transformed from continuous value)	Relative vibrational motion; indicates the mobility of each C□ atom. The mutation occurs at a highly dynamic (or highly rigid) residue if B-factor$_{Norm}$ ≥ 97.5 (or ≤ 2.5) percentile.	PredyFlexy [75]
	RMSF$_{Norm}$	Categorical (transformed from continuous value)	Root-mean-square fluctuation; indicates the movement amplitude of each Cα atom over a period of time. The mutation occurs at a highly dynamic (or highly rigid) residue if RMSF$_{Norm}$ ≥ 97.5 (or ≤ 2.5) percentile.	PredyFlexy [75]
Protein flexibility (large amplitude)	FlexPred label	Categorical	The mutation occurs at a predicted conformationally rigid or flexible site.	FlexPred [76, 77]
	Hinge site	Categorical	The mutation occurs at a predicted hinge site.	FlexServe [159]
	Binding site	Categorical	The mutation occurs at a predicted binding site.	3DLigandSite [72]
Drug binding capability	RSA	Categorical (transformed from continuous value)	Relative solvent accessibility; indicates percentage of solvent accessible compared to the fully exposed peptide. The mutation occurs at a buried site if RSA ≤ 20%, and at the protein core if RSA ≤5%.	PoPMusic 2.1 [120]
	Γ	Categorical (transformed from continuous value)	Degree of residue non-optimality (summation of all stabilizing $\Delta\Delta$Gs). A mutation occurs at a highly non-optimal residue if $\Gamma \leq 5$ percentiles (an indication of being a catalytic site).	PoPMusic 2.1 [120]

Table 5.3 (continued)

Structural feature	Parameter	Variable type	Description	Tool
Protein-protein interaction	Protein patch	Categorical	The mutation occurs at a conserved cluster on a protein surface (an indication for site of intermolecular interactions).	PatchFinder [73, 74]
Residue localization	Protein domain	Categorical	The mutation occurs at a conserved protein domain family.	InterPro [155]
	Structural site	Categorical	The mutation occurs at a structural site (any residues excluding the one with the following annotations: molecular processing, amino acid modifications, sequence motif, or sequence bias).	Uniprot [150]
Amino acid dissimilarity	Grantham score	Categorical (transformed from continuous value)	Indicates the dissimilarity of an amino acid pair. The substitution is radical (or conservative) if Grantham score is ≥ 100 (or < 100).	Grantham matrix [56]
	Physicochemical properties of amino acid side chains	Categorical	Indicates the classification of hydropathy index (hydrophobic and hydrophilic properties), volume, charge, and hydrogen donor/acceptor for amino acid side chains.	IMGT standardized criteria for amino acid side chain [156]
Amino acid secondary structure preference	Preferred / break	categorical	Indicates propensities of amino acids towards a particular secondary structure.	[157]and PoPMusic 2.1 [120]

Salt bridge is the most commonly observed noncovalent interaction that contributes to the stability of entropically unfavorable protein folds [160]. The cut-off distance to define salt bridges in a protein structure varies, but generally centers around the distance of 3.0 Å that is used to define hydrogen bonds (H-bonds). A distinction between salt bridge and long-range electrostatic interaction between charged groups was made [161], such that salt bridges are strong H-bonds between donor and acceptor atoms within the distance cut-off of 2.8 Å, whereas long-range electrostatic interactions can have variable cut-offs since the association energy depends upon types of environmental media. We used the WHAT IF server [158] to characterize bonding patterns between opposite charged protein residues. We followed the default distance cut-off of 7.0 Å so that the predictions were suitable for both salt bridges and long-range electrostatic interactions.

Hinge residues are typically located at inter-domain boundaries of a protein. Hinges allow the domains to move relative to one another upon binding to another molecule, or upon activation/deactivation of the protein. Hinge points do not involve large number of residues and each region can rotate at a very large degree. All of the proteins in our dataset occur as multi-domain proteins, therefore, the study of hinge regions is considered very informative since their cooperative motions are a critical element of protein-drug or protein-protein interactions. There are numerous computational approaches for predicting this type of large-scale protein movement; their computational resource requirements also vary over dynamic time scales. One prominent tool, FlexServe [159], detects hinge protein residues using a coarse-grained based method as an alternative to the resource-intensive Molecular Dynamic (MD) simulations. We performed hinge point predictions under the normal mode analysis with the following default settings: Kovacs algorithm

(distance-dependent potential), inter-atomic force constant 40 kcal/mol*Å, and maximum distance cut-off for close atomic pairs of 3.0 Å.

Statistical comparison of disrupted protein residues and the implementation of SDS Pharmacogene for predicting the structural significance of a variant

Our systematic 3D screening aims to identify protein characteristics that are commonly disrupted by functional variants. The specific details for performing statistical comparisons of disrupted protein residues can be found in our previous work [110]. Briefly, after all residues in the 45 protein structures had been evaluated by each of the selected tools, we populated all measures/predictions and generated a distribution of the scores (if the tool outputs a continuous parameter; i.e., B-factor, RMSF, or Γ) (**Table 5.3**). Lower- and upper-bound thresholds were defined as suggested in the literature or empirically. Empirically-defined thresholds were set at the top and bottom 2.5 percentiles of each distribution. Based on the selected cut-offs, we transformed all measures into a 2-level category: extremely small and extremely large values. Variants whose measures or predictions go beyond the defined threshold are most likely to be aberrant residues (with disrupted protein structures). We repeated the three steps of score assessment, threshold determination, and category assignment to all of the five continuous measures used in our study. For tools that report the prediction in categorical variables; e.g. predictions of disulfide bonds, salt bridges and electrostatic interactions, and stabilization centers, etc., we counted the number of variants that fall into each category and reported their absolute numbers.

Fisher's exact test was used to determine whether the proportions of functional and neutral variants for a particular structural feature are significantly different. We applied this statistical method to variants located in our protein 3D structures (371 functional and 1,811 neutral mutations) and report the test statistics when p-values are significant (α = .05). Structural characteristics that harbor different proportions of functional vs. neutral variants were identified (n= 11 features).

Further refinement of the selected structural features was then performed according to the proportion comparisons (presence/absence) of the disrupted characters among the three groups of variants. Five out of eleven indicators showed significant differentiation of PGx variants from neutral mutations; in each case they also suggest PGx variants to have similar structural properties to the one that altered the protein functionality (functional variants) (**Table 5.4**). A list of five strong suggestive characters was populated as a "Structural Disruption Index". Given the predictions of the five suggestive characters, we calculated the consensus prediction for the structural significance of a variant: "positive" if a variant induces/interrupts any of the five suggestive characters, and "negative" otherwise.

Table 5.4: Fisher's exact test statistics for enriched structural features present in functional mutations and the selection for strongest predictive features for structural disturbance (Structural Disruption Index). The complete statistics, including the absolute number of functional and neutral variants in all categories of structural indicators can be found in **Supplementary Table 5**.

Structural features	Indicators	Descriptions	Preliminary selection of structural features		Final selection of structural features as Structural Disruption Index (SDI)	
			Fisher's exact test (one-tailed)	n	Likelihood Ratio (Chi-square)	% of functional/neutral/PGx variants (n=2214)
Protein stability	Is a stabilizing residue (SR)	Wild type residue = a stabilizing residue (SR)	0.0037**	2182	.0154*	(2.77/0.79/1.39)
	Induced any changes of stability	Mutant residue reduces or increases stability ($\Delta\Delta G \geq 0.5$ or ≤ -0.5 kcal/mol)	0.0311*	2182	-	-
Protein flexibility	Conformationally rigid	Located at conformationally rigid site (FlexPred label = rigid)	0.0218*	2182	-	-
Drug binding capability	At binding site	Located at binding site	0.0003***	2182	-	-
	At 10 Å of binding site	Located within 10 Å of binding site (Cα-Cα distance)	0.0005***	2182	.0058**	(33.24/25.49/33.33)
Protein-protein interaction	At patch	Located on protein patch	< 0.0001*	2182		
Residue localization	At core	Located at the core region (RSA ≤ 5%)	0.0021**	2182	.0151*	(35.73/29.14/38.89)
	Induced Gly/Pro change	Induced Gly/Pro change	0.0015**	2157		

Table 5.4 (continued)

Structural features	Indicators	Descriptions	Preliminary selection of structural features		Final selection of structural features as Structural Disruption Index (SDI)	
			Fisher's exact test (one-tailed)	n	Likelihood Ratio (Chi-square)	% of functional/neutral/PGx variants (n=2214)
Amino acid dissimilarity	Large AA dissimilarity at structural site	Induced large amino acid change (Grantham score ≥ 100) when located at structural sites‡	< 0.0001***	3165	< .0001* **	(35.84/22.98/36.11)
	Change of volume, any site	Induced volume change (very large --> small/very small, very small --> large/very large) when located at any sites‡	0.0002***	3316	-	-
	Worse hydropathy at buried site	Induced unfavorable hydropathy (hydrophobic to hydrophillic) when located at buried site	0.01**	1096	< .0001* **	(4.43/1.74/16.67)

Significant p-values (α =.0001, .01 and .05) are designated with '***', '**', and '*', respectively.

During the preliminary selection of structural features, except for the two tests marked with '‡', the dataset include variants whose 3D structures are available (371 functional mutations and 1811 neutral mutations in 45 gene products). The tests marked with '‡' were performed on protein sequences; the dataset represents the maximum of 3316 data points (779 and 2537 functional and neutral mutations, respectively). During the final selection of structural features as Structural Disruption Index (SDI), all tests were performed on non-overlap set of functional, neutral and PGx variants (n= 361, 1781, and 72 variants, respectively). The percentages of variants with particular disturbance features are reported.

In addition to the binary classification of structural importance, our implementation also provides a composite structural score of those five measures—"Structural Disruption Score" (SDS), the score ranges from 0 to 5 (each point is taken from an individual – "Structural Disruption Index" (SDI)). Since a structural feature indicates only one function, among possible multiple roles a protein residue may have. Therefore, mutating amino acid residues with multiple hits for potential structural disturbance should be of high priority; this interpretation is represented by the higher SDS (closer to 5).

Evaluation of conventional deleterious classifiers and protein characters for assessing amino acid variants in pharmacogenes

The pharmacogenomics (PGx) variants represent an independent set of genomic data that were used to evaluate the performances of sequence conservation-based predictors, protein characters, and SDS for the deleterious evaluation of amino acid variants in the 48 VIPs. Some PGx variants were found to overlap with the training set, therefore, 12 functional mutations and 32 neutral mutations that are duplicates with PGx variants were excluded from all evaluation tests. An additional filter was used to remove all variant with no available structural data, leaving a unique and non-overlapping set of 361 functional variants, 1781 neutral variants and 72 PGx variants. We used this dataset to performed several evaluations.

The first test compared means (ANOVA tests) and distributions of five continuous structural parameters (P(Flexible), RSA, $\Delta\Delta G$, Γ, B-factor, and RMSF) (**Table 5.3**) and six continuous scores of conventional deleterious classifiers (SIFT [16],

PolyPhen2_HumDiv and PolyPhen2_HumVar [19], LRT [17], MutationTaster [20], and MutationAssessor [18]) across three types of amino acid mutations.

The second evaluation, chi-square test with $\alpha = .05$, was performed on the numbers of variants with binary consensus structural disturbance predictions (positive/negative), as well as the categorical levels of structural disturbance (score 0 to 5). The results were compared with similar tests on the consensus deleterious predictions (whether or not $\geq$ 3/6 conservation-based tools assign the variant to be "deleterious") and the categorical assignment "deleterious count" (the number of tools that predict a variant to be deleterious, maximum score = 6).

Next, we focused on the classification of PGx variants to match the annotation of PharmGKB's three levels of annotation confidence. We evaluated the predictive power of different methods for PGx variants whose 3D structures are available (n=386 associated variants, equivalent to 72 unique variants). To test whether standard conservation-based predictors can statistically distinguish variants of each level, we first clustered the 386 variants into 3 groups: level 1A/1B, level 2A, and level 3. Next, we compared the average deleterious prediction scores from the six conservation-based predictors. The comparisons were performed using each paired Student's t-test on rescaled prediction scores (range 0 to 1). The tests were carried out with both the original and the revised listing of pharmacogenomic associations. The original list comprises of 386 associations with structure data.

Results and discussion

Location of missense variants in protein structures

A total of 113 conserved protein domain families harbor missense variants in our genomic dataset. To further define the relative abundance of the 3,316 amino acid mutations in specific protein domain families, we compared the normalized variant percentages between the functional and neutral variant groups. Although many protein domain families have both types of variant, we found that some domains contain much higher percentages of variants of either group. Of all 113 protein domain families that are present in the 48 proteins, we identified 36 domains with greater abundance of functional variants, 42 domains with greater abundance of neutral variants, and 35 domains with only neutral variants (**Figure 5.1, Supplementary Table B.4**). Two explanations contribute to these observations. First, there is a strong preference for some protein domains to harbor more or less functional variants, since the relative variant abundances within the first two domain types are significantly different ($p < .0001$, Student's t-test, df 76). Second, the analysis discovered some protein regions with insufficient sampling depth from current sequencing technology.

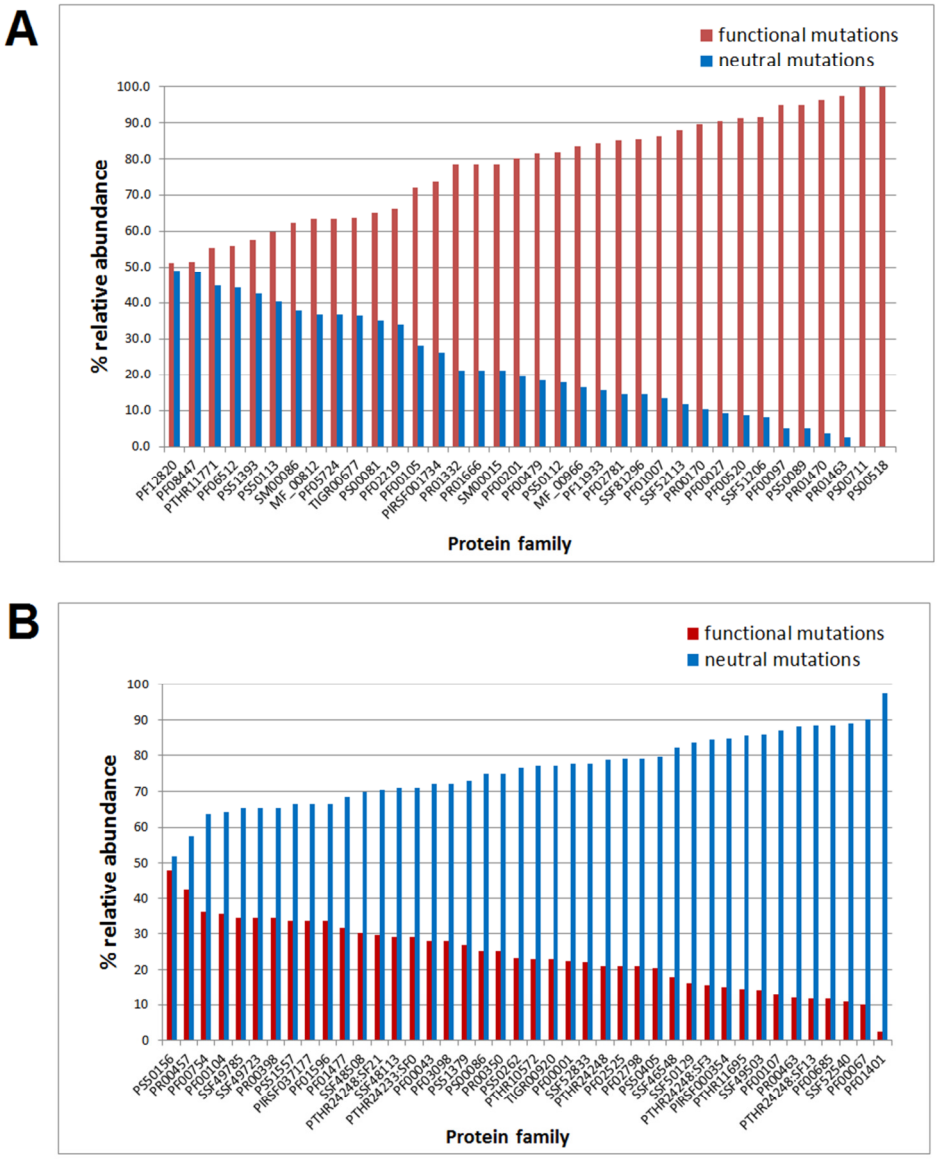


Figure 5.1: Relative abundance of functional and neutral variants across conserved protein domain families. Of all 113 protein domains that are present in the 48 proteins, 36 domains harbor more functional variants (**Figure 5.1A**) and 42 domains contain more neutral mutations (**Figure 5.1B**). The remaining 35 protein domains only have genomic data for neutral variants. The complete list of domain names and relative abundances of functional and neutral variants is provided in **Supplementary Table B.4**.

Physicochemical change of amino acids at different protein regions

We used the Grantham matrix [56] to measure physicochemical differences of each amino acid substitution. We compared the proportion of conservative vs. radical mutations of functional variants with respect to the variant localization along the protein chain (**Supplementary Table B.5**). All analyses suggest that only when functional variants occur in non-structural sites (i.e., as propeptide, signal peptide, sequence motif, zinc finger, repeat, or modified residue), do they not tend to associate with large dissimilarities. For other regions, functional variants always have a stronger preference towards substitutions with large amino acid dissimilarity.

The first comparison examined Grantham scores of variants in structural sites relative to non-structural sites. As expected, when functional variants are located in structural sites of the protein, they tend to be more of radical type (Grantham scores ≥ 100) ($p < .0001$). On the other hand, functional variants do not seem to induce large amino acid changes when the altered protein residues locate elsewhere ($p = .6145$).

Next, we were interested to see if the localization of variants in conserved protein domains would increase the likelihood of drastic amino acid change, or if functional variants have a tendency to induce radical substitutions throughout the protein chain. Comparing amino acid substitution scores (Grantham ≥ 100 vs. < 100) of variants in protein domains to ones located at domain boundaries did not suggest unequal proportions ($p = .42$, one-tailed Fisher's exact test, n=3,316). Conversely, when functional variants are situated at either protein domains or at protein boundaries, they

tend to equally promote large amino acid change compared to neutral mutations ($p < .0001$ and $p = .0045$ for the two locations, respectively).

Lastly, we tested whether predicted binding sites tend to harbor more functional variants with a Grantham score ≥ 100. The test statistics indicate functional variants do tend to associate with amino acid dissimilarity when they are predicted to alter binding sites ($p = .0259$, n=145), and this tendency is much stronger if they are located within 10 Å of the binding site ($p < .0001$, n=591).

Additional tests for unfavorable amino acid changes (hydropathy, volume, charge, hydrogen donor/acceptor) indicated that functional variants do not directly have large physicochemical dissimilarity at/around the binding sites ($\alpha = .05$), and that change in side chain volume is the only significant physical disturbance caused by functional variants ($p = .0002$ at any sites, n=3,316; $p = .0046$ at around 10 Å of binding sites, n=591) (**Supplementary Table B.5**). The three most insignificant indicators are ligand-specific: 40/48 VIPs serve as drug targets for 530 drugs and 23 genes have known pharmacological actions for 270 drugs [152]. This large number of highly interactive genes and the diversity of their drug partners may explain why we did not observe any enrichment in structural disturbances of the three physical changes. Besides the potential disruption in binding cavity volume, other notable changes we detected from functional mutations is that they tend to alter the hydrophobicity of buried residues ($p = .01$, n=1,096).

The analysis of secondary structure preference for amino acids pointed out that although functional variants tend to have gain or loss of glycine or proline residues (two unusual

amino acids that are either very conformationally flexible or rigid, respectively), the secondary structure of the protein is still well maintained. We did not observe any structure break in the six secondary structure types (coil, strand, 3-turn helix, α-helix, bend, and turn) (**Supplementary Table B.5**).

Implementation of systematic 3D screening for structure-function relationships of amino acid mutations

PharmGKB is a knowledge-based database that provides curated resources on genetic variation and drug responses [146]. As of February 2014, there are 249 peer-reviewed scientific articles that were written by PharmGKB affiliated researchers. Among these, 25 articles address the pharmacogenomics of an individual gene, and 29 publications emphasize drug metabolism pathways. The remaining 195 articles mostly describe population studies or dosage studies, or are review articles. Most significantly, only one article [35] utilizes 3D structure data (stability change) to improve the prediction accuracy of a popular conservation-based deleterious predictor—PolyPhen2 [19]. Their study showed that with a single structural parameter added to the sequence data, the deleterious prediction performance improved from 82 to 85% accuracy, and the Matthew's correlation coefficient (a quality estimator for binary classifications) increased from 0.63 to 0.70. Although the authors of PolyPhen2 had mentioned that their algorithm includes 11 predictive features (6 conservation-based, 4 structure-based, and 1 derived parameter), we reasoned that since the number of included structural data annotations was very limited (only ~10% of the training set have available structural data), it can be further improved upon. Note that the four structural features included in PolyPhen2 are

side chain volume, accessible surface area, accessible surface area for buried residues, and B-factors (indicate mobility of individual atoms) [19].

A recent review presenting a protein structural perspective on drug responses [162] summarized the different structural characteristics that are important for pharmacodynamic and pharmacokinetic mechanisms. The authors considered five major classes of computational tool for examining drug-protein relationships: structure visualization tools, structure prediction tools, binding site prediction and comparison tools, ligand docking and scoring tools, and MD (molecular dynamics) simulation tools. We added to this list some other potential protein analyses that incorporate other types of interactions that occur within or between proteins even when no drug molecules are bound (**Table 5.3**).

The nearly complete coverage for functional and neutral mutations which can be mapped to structural data of the 45 VIPs (average variant coverage of 85% per gene, range 75 to 95%, 95% CI, **Supplementary Table B.3**) enhances the ability to study the structure-function relationships of missense variants. We employed a similar approach in our study of epilepsy-associated variants [110] where we showed that quantifying the structural disruption features caused by functional mutations can prioritize risk variants. Furthermore, we added a few explicit test categories in this current study, namely, inter-residue bonding, residue localization, amino acid dissimilarity, and secondary structure preference of amino acids.

Implementation of SDS Pharmacogenes for predicting structural significance of a variant

The analysis of context-dependent mutation effects describes a list of physical protein features that are commonly found in functional mutations as oppose to in neutral variants. These enriched characters are important because they suggest specific protein attributes and/or residues that are less tolerant of disruption. We have integrated scores of five selected components in "Structural Disruption Index" (SDI) into a composite "Structural Disruption Score for Pharmacogenes" (SDS Pharmacogenes). The score provides a measure of the likelihood that a mutation is damaging. It can be used to evaluate all missense variants found in this gene set whose protein 3D structure is available (currently 45/48 gene products) and thus to derive a list of mutations that may cause altered protein functions when interacting with drug molecules, contributing to inter-patient variability of drug responses.

Among the 68 structural features we compared between functional and neutral variants, 11 features were found to be more common in functional mutations (**Table 5.4**). Six indicators may have pharmacodynamic effects due to altered protein stability, binding sites, or protein conformations. One indicator, conformational rigidity, also plays an important role in pharmacokinetic mechanisms of many proteins, including the CYP450 system [163]. Two indicators also have close relationships with sequence conservation, namely, stabilizing residues and protein patches. In addition, all interior protein residues (RSA $\leq$ 5%) appear to be highly conserved. The full list of structural characteristics that were tested during the algorithm development and the correspondent test statistics are provided in **Supplementary Table B.5**.

Next, we aimed to narrow the list of enriched structural features to a small number of protein characters that best assign PGx variants to resemble functional mutations, rather than to the neutral ones. **Table 5.4** indicates that eight of the eleven structural measures illustrate a shift in PGx variants in the same direction as functional mutations, and away from the neutral ones (the three exceptions are induced Gly/Pro change, at patch, and at binding site). Among the eight features, five have significant statistics that cluster functional and PGx variants to the exclusion of neutral ones; the features are: unfavorable hydropathy change at buried site, large amino acid dissimilarity (Grantham score ≥ 100) if located at structural sites, located at core region (RSA $\leq 20\%$), located within 10 Å of predicted binding sites, and is a stabilization residue (SR). Jointly they define a "Structural Disruption Index" (SDI) that can be used to infer structural (and maybe functional) significance of an amino acid variants based on the consensus prediction of structural disturbance (positive/negative) and the magnitude of structural impact (SDS).

Structural characteristics of functional variants

The five components of the "Structural Disruption Index" (SDI) are unfavorable hydropathy change at buried site (RSA $\leq 20\%$), large amino acid dissimilarity (Grantham score ≥ 100) if located at structural sites, located at core region (RSA $\leq 5\%$), located within 10 Å of predicted binding sites, and is a stabilization residue (SR).

Hydropathy is an important property of amino acid side chains. Hydrophobic residues tend to be at buried sites (RSA $\leq 20\%$) while amino acids with hydrophilic side chains are more commonly found on protein surfaces. Side chains with similar hydropathy attract each other. The clustering of hydrophobic side chains within the protein core is a

major force that holds a protein structure in place [161]. Any amino acid substitutions that reduce hydrophobic interactions may cause serious folding defects.

Grantham scores serve as a general criterion to classify amino acid dissimilarity for a pair of amino acids. The score is derived from an index that best correlated physicochemical properties of protein residue to the substitution frequencies [56]. The properties include amino acid composition, polarity, and molecular volume. We did not detect any difference in the proportion of large amino acid replacements (Grantham score ≥ 100) among the three levels of PGx variants. However, large Grantham score is a good indicator for separating PGx from neutral variants ($p < .0095$, Fisher's exact test).

Protein residues that have significant roles can be either solvent-accessible (e.g. some ligand binding sites or protein-protein interaction sites), or can be solvent-inaccessible (e.g. several ligand binding sites, catalytic sites). In addition, conserved residues within the protein are likely to have important structural roles in maintaining the correct protein fold. This feature is described as "stabilizing residue" (SR) [71], which indicates the protein residue has: (1) high surrounding hydrophobicity (sum of hydrophobic indices of all its neighbor residues within 8 Å is greater than 20 kcal mol^{-1}), (2) high long-range order (the percentage of long-range neighbors ($\geq$ 12amino acids apart) with the C_α-C_α distance of ≤ 8 Å is greater than 2% of the total number of protein residues), and (3) high conservation score (score $\geq 6/9$ based on ConSurf predictions [164]).

Although we only detect 1% (215/21457) of all the protein residues in our 3D structure data of 45 proteins to be qualified as SR (the number can be increase if we relax any of

the three criteria), the proportion is higher (p = .0034, Fisher's exact test) in functional variants (2.8%, 10/361 residues) compared to neutral variants (0.8%, 14/1781 residues).

Among the72 PGx variants, only one mutation is detected as an SR (equivalent to 1.4% of the dataset). This PGx variant, Y240C in *TPMT* (rs1142345) is classified as*TPMT*3* allele and is associated to five drugs at the PharmGKB's level 1 annotation (azathioprine, mercaptopurine, purine analogues, and thioguanine) and one drug (cisplatin) in PGx level 3. Four conservation based predictors (except MutaionTaster and MutationAssessor) predict the variant to be deleterious.

In addition to being predicted as an SR, this Tyr240Cys substitution is independently predicted by by PoPMuSiC software [120] to severely reduce protein stability (ΔΔG 2.24 kcal mol^{-1}). The functional significance of this variant has been confirmed by experimental studies [146] which demonstrate the loss of several side chain contacts, accelerate in vitro degradation, and reduce the protein activity. In addition, an MD simulation study indicates this variant of *TPMT* causes the most structural deformation during simulations [165].

An exceptional case

In general, the 386 pharmacogenomic-associated variants harbor at least one type of the eleven structural disrupted features (maximum SDS = 3/5). Only one variant, R26H in *CYP2D6,* has 0 hits in our SDI (SDS = 0/5) and in fact 0 hits among the initial selection of 11 predictive features. Prediction results from 5/6 conservation based predictors indicate the variant is benign (prediction scores are at the extreme end for "benign").

MutationTaster [20] is the only program which predicts this amino acid change to be deleterious with a score of 0.999 (cutoff = 0.5).

The R26H amino acid change of CYP2D6 is identified as the CYP2D6*46 allele (rs28371696) [166]. CYP2D6 metabolizes a wide range of drugs—up to 25% of commonly used prescriptions, such as, antidepressants, antiarrhythmic agents, neuroleptics, opioids, and antihistamines [167]. R26H in *CYP2D6* has been linked to 31 drugs in the PharmGKB's pharmacogenomic association schema levels 1-3. Overall, Africans metabolize CYP2D6 substrates at a slower rate than Caucasians owing to the higher alternate allele frequency (1.68% vs. 0.02%) [151].

Of all 68 structural features that were tested during the development of SDS pharmacogenes (**Supplementary Table B.5**), the R26H does not unusual for any of those features. Arg26 is locate at the exterior loop of the protein (RSA = 53%). The amino acid change does not induce a secondary structure break at this coil region. In addition, Arg and His have very comparable physicochemical properties (Grantham score = 29). The substitution does not induce any stability change ($\Delta\Delta G$ = -0.09 kcal mol^{-1}). The only structural feature that comes close to being positive for this variant is the fact that Arg26 is predicted to be a non-optimal residue (several of its mutations will indeed increase the protein stability) The Γ parameter (designates the degree of sequence non-optimality) for Arg26 is -3.16 kcal mol^{-1} (positive cut-off is when $\Gamma \leq 3.47$ kcal mol^{-1}; an indication for resemble a catalytic residue).

CYP450 enzymes consist of an N-terminal transmembrane α-helix and a catalytic domain. While the structure of the catalytic domain is well-resolved, the N-terminal

domain cannot be elucidated due to insufficient x-ray density. Four crystal structures of human CYP2D6 are currently available [60], but the first 33 residues (N-terminal) have been replaced with an 11 amino acids long peptide to improve crystallography. The full length model of CYP2D6 predicts reside 1-25 as the N-terminal transmembrane α-helix, and residues 26-33 to be the linker between the N-terminal and the catalytic domain. Molecular dynamic simulations of related enzymes using full length protein structures revealed the N-terminal region exhibits a large range of flexibility [168, 169] and the presence of N-terminal region alters the conformational dynamics of opening and closing ligand channels compared to the N-terminal truncated crystal structure of CYP2D6 [170]. Further assessments are required for the functional roles of Arg26 in this enzyme.

Performance of conservation-based predictors and SDS for assessing amino acid variants in pharmacogenes

1) *Classification of functional, neutral, and PGx variants*

The six deleterious predictors we chose for comparison with the SDS Pharmacogenes score differ in the underlying methodologies. Three algorithms (SIFT [16], LRT [17], and MutationAssessor [18]) were built from evolutionary conservation, while the other three (MutataionTaster [20], and PolyPhen2_HDIV, PolyPhen2_HVAR [19]) incorporate various SNP data as predictive features. SIFT calculates the normalized probability of a substitution from sequence alignments of orthologous proteins [16]. LRT uses sequence alignments of related proteins from vertebrate species to check for negative selection of codons [17]. MutationAssessor partitions sequence alignments to identify evolutionary conserved sites that contribute to protein function [18]. MutationTaster combines several

sources of information to generate a supervised-learning classifier [20]. PolyPhen2_HDIV and PolyPhen2_HVAR use protein sequence (orthologs and paralogs) and structure-based information (side chain volumes, accessible surface areas, accessible surface areas for buried residues, and B-factors) to generate a Bayesian model for variant deleteriousness [19].

We found the classification of genetic variants (functional, neutral and PGx) in the 48 pharmacogenes is challenging. The poor performance of most conservation-based scores is illustrated by the plots in **Figure 5.2.** Although the consensus prediction, deleterious count, for the six tools can rank functional and neutral variants in the correct order, the deleterious assignment for PGx mutations is not meaningful. Likewise, each pair of Student's t-tests indicates all programs properly rank the functional variants to be more deleterious than the neutral ones ($p < .0001$ for all predictors). On the contrary, every program failed to place PGx variants closer to the functional mutations. More importantly, PGx variants were assigned less deleterious scores (less damaging) than neutral variations ($p < .0001$ for all predictors except $p=.0112$ for SIFT). Density plots suggest the conservation-based programs have a tendency to under-predict the damaging effects of variants in this set of 48 pharmacogenes.

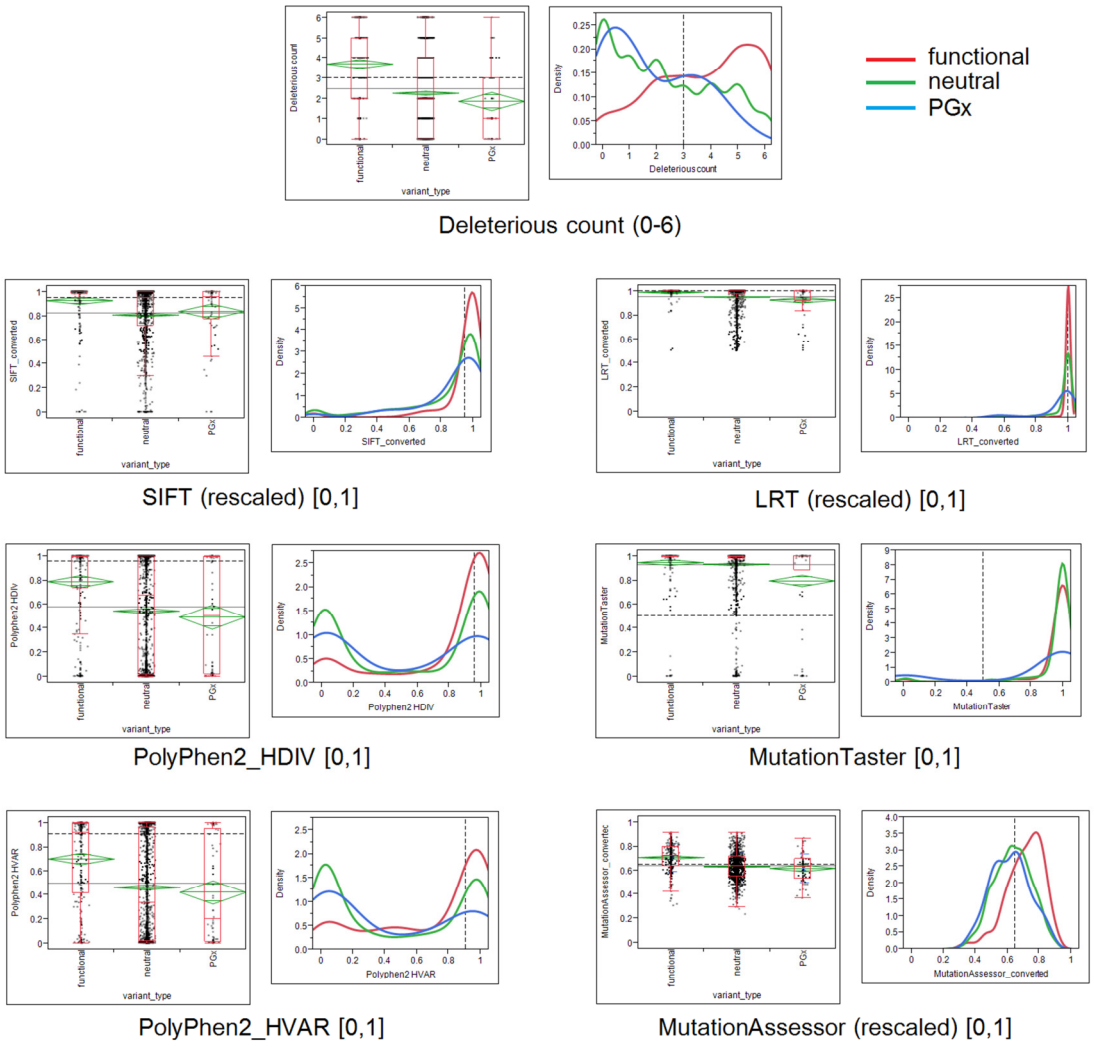


Figure 5.2: Performance of conservation-based predictors across the three types of mutations in 48 VIPs. The first row represents the consensus prediction for deleteriousness (deleterious count, score 0-6). The remaining rows demonstrate the score distributions for different types of variants (n= 361, 1781, and 72 for function, neutral, and PGx variants, respectively). Some scores were rescaled to [0, 1] so that deleterious scores of the six predictors can be interpreting in the same manner—scores closer to zero indicate the prediction is "benign" while score closer to one indicate "deleterious". The horizontal (vertical) dashed black lines in the scatter (density) plots indicate the deleterious threshold for each predictor.

We performed chi-square tests on the numbers of variants that receive consensus "positive" or "negative" predictions for inducing any structural disturbance. The test statistics demonstrated that the proportions of "positive" variants differed significantly among the three groups of mutations ($p < .0001$). Approximately 78% of PGx variants (n = 56/72) will interrupt at least one of the five elements of the SDI, and this bias is in the same direction as the 66% of functional mutations (n=239/361), whereas only 57% of neutral mutations (n=1013/1781) have a positive SDI (**Figure 5.3**). Structurally disrupted variants are more abundant in PGx than neutral variants ($p = .0002$, n = 1,853, Fisher's exact test), and in slightly more excess in PGx than functional variants ($p = .0647$, n = 433, Fisher's exact test).

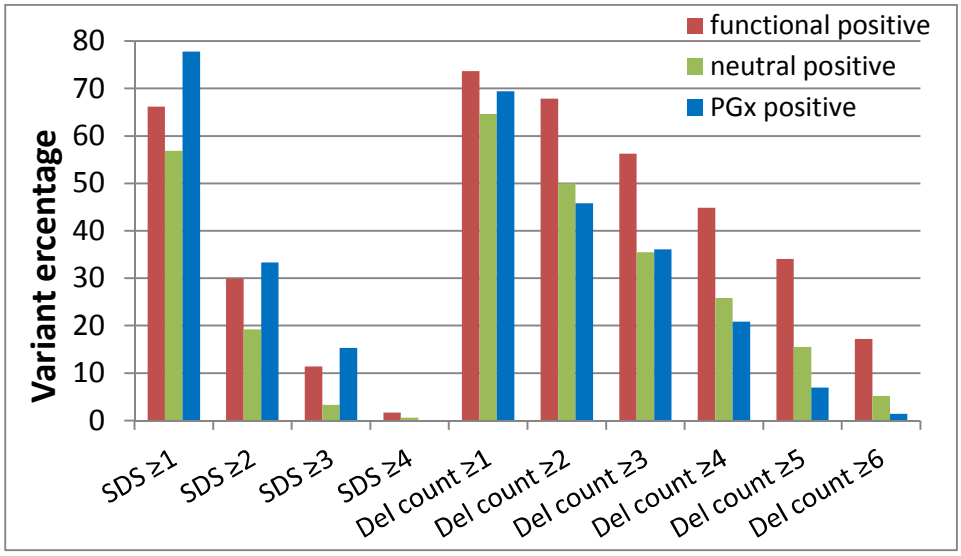


Figure 5.3: Percentages of variants with respect to their scores for structural significance (Structural Disruption Score; SDS) and the consensus scores for conservation-based deleteriousness (Deleterious count). The graph shows the percentage of positive variants (variants that are predicted to disrupt at least one of the five components of Structural Disruption Index (SDI)) which have score of at least *n* for SDS and deleterious count. Note the SDS can distinguish between PGx and neutral variants, and PGx variants receive comparable or higher SDSs than functional variants in all SDS thresholds, whereas deleterious count (del count) assigns the score for PGx and neutral variants with no meaningful interpretation.

We also observed that SDS correlates highly with the deleterious count (p = < .0001, n=1,902), although the correlation is not strong (Pearson's correlation coefficient = .248). SDS also correlates (p < .0001) with SIFT, PolyPhen2_HDIV, PolyPhen2_HVAR, and MutationAssessor but with small coefficient (range 0.18 to 0.29).

2) *Classification of three levels of PGx variants*

PharmGKB has annotated 776 variant-drug pharmacogenomic associations for the 48 VIPs (**Table 5.1**).The association strength is classified into four levels: 1A/1B, 2A/2B, 3, and 4. The most number of associations are linked to CYP450 enzymes; their variants account for 83, 81 and 48% of all levels 1A/1B, 2A/2B, and 3 variants, respectively. In addition, we observed that missense mutations contribute approximately 61% (471/776) of the overall associations, consistent with a recent remark that states 60% of disease-causing variations are missense variants [14]. Among the remaining associations, just 3% (22/776) are caused by nonsynonymous mutations, and 36% (282/776) are induced by non-coding variants (variants in intergenic, intron, or 3'UTR regions). Their associations with pharmacological traits are not strong and the variants are mostly classified as level 4 in the PhamGKB's schema. Among the 471 missense mutations, many have pharmacogenomic associations with more than one drug (ranges 1 to 10), but the levels of association vary upon the interacting compounds. The CYP2D6 gene has the maximum number of FDA-approved pharmacogenomic drug labels per variant (seven drugs: amitriptyline, clomipramine, desipramine, doxepin, imipramine, nortriptyline, and trimipramine) (http://www.fda.gov/drugs) (accessed March 7, 2014).

The 471 variant-drug pairs uniquely involve 81 missense mutations (levels 1-3) of 27 proteins. After we applied the six conservation-based deleterious predictors to the 471 missense variants of all annotation levels, we compared the average prediction scores across the three levels: 1A/1B, 2A, and 3. Student's t-test statistics indicate only one program, SIFT [16], can differentiate variants of the three pharmacogenomics association levels based on the score averages ($P_{level\ 1\text{-}2} = .0042$, $P_{level\ 1\text{-}3} = .0015$). The trend for SIFT score is as expected; SIFT predicts pharmacogenomic-associated variants in level 1A/1B and 3 as the most and the least deleterious groups, respectively. This observation is consistent with PhamGKB's curated levels of associations.

Results from different conservation-based tools also demonstrate the unforeseen difficulties in analyzing genomic variants in this gene set. The ability of SIFT [16] and inability of PolyPhen2 [19] to cluster PGx variants is worth investigating, since both programs perform exceptionally well in general. The poor performance of PolyPhen2_HDIV is surprising, since the structural information used by PolyPhen2_HDIV is related to the enriched disrupted features we detected among functional variants (**Table 5.4**). More importantly, the training set PolyPhen2_HDIV used comprises of SNPs that cause Mendelian diseases by affecting protein stability and function [19], therefore, the substantial contribution of structural features is expected.

SDS is a numerical parameter; the value ranges from 0 to the maximum possible SDS (currently 5). In the dataset of 386 pharmacogenomic associations, the average SDS is 1.26 and the maximum SDS is 3. We did not detect any differences between average SDS among the three levels of PGx associations; the observation is in agreement with the deleterious count (representing the number of conservation based tools that predict a

variant to be deleterious, average score 1.80/6). A major reason is that the dataset of 386 associated variants are very much redundant (uniquely mapped to 72 variants). The identical protein residue can interact with many small molecules and other proteins in through various modes of interactions.

Limitations of pharmacogenomics studies

The need for expanded application of sequencing technologies in pharmacogenetics, especially for the ability to uniformly and fully capture all genomic variations within genes, was raised by [23]. They assessed the polymorphism coverage of 253 pharmacogenes from the 1000 Genomes Project [171], and discovered that no current genotyping technologies cover more than 85% of residues within each gene. More importantly, data for only 30% of missense mutations are being generated across all sequencing platforms. This notable remark was also observed in our genomic dataset, in which we detect some protein domains seriously lack the data for functional variants (**Figure 5.1**).

In addition to more extensive databases of variation in pharmacogenetic loci, exhaustive mutagenesis data is another key to the generation of accurate predictive tools for damaging variants. At this stage of genome analysis, the data on human studies is used to deduce the functional impact of variants in a binary format (functional vs. neutral variants). Experimental data provides observable phenotypes with measurable effects. The information has the potential to greatly enhance the development of deleterious prediction algorithms. The data will provide a means to guide, refine and validate prediction models.

A future direction of this work is to explore mutagenesis data and/or genomic data of the pharmacogenes to aid the refinement of feature selection. We also plan to populate SDSs for all amino acid mutations in the currently available set of 3D structures (n=45). Our future implementation of a database and a web-application for SDS Pharmacogenes will allow an easy retrieval of the pre-computed scores, thus promoting the practical utility of this research.

Conclusions

Our implementation of "SDS Pharmacogenes" provides a foundation for investigating the roles of amino acid mutations in drug responses from the standpoint of structural analysis. The pipeline introduces several conceptual shifts in assessment of genomic variations. First, the evaluation is strictly structure-based, therefore, offers a unique opportunity to explore the effects of amino acid mutations within the structure environment. Second, the analysis is systematic, through the investigation of structure-function relationships in the context of each protein topology. Third, the implementation is scalable, suitable for the large and growing accumulation of genome data. Fourth, the prediction is informative, because the SDS Pharmacogenes combines results from several analyses, and the consensus prediction for structural disturbance (positive/negative) is meaningful. Unlike existing deleterious scores that do not consistently classify Pharmacogenetic VIP mutations as functional, we have identified five protein structure features that are significantly enriched in VIP variants. Further refinement of SDS Pharmacogenes, including the database-driven web application, will permit anyone, with or without protein expertise, to quickly obtain the comprehensive predicted drug responses for the variant of interest.

Acknowledgments

This work was supported by start-up funds from the Georgia Tech Research Foundation to GG, and TP was supported by the School of Biology at Georgia Tech.

CHAPTER 6: CONCLUSIONS

Sequencing technology development and the outlook for genomic variant analysis pipelines

The rapid technological advances in sequencing technology are promising ever more breakthroughs from genome studies. The cost of sequencing is now estimated at about $1000 per genome, a huge drop compared to the first completed genome data 10 years ago that cost nearly $3 billion [26]. *HiSeq X Ten*, the newest sequencing platform from Illumina, is advertised to be capable of sequencing 18,000 genomes per year, while keeping the cost close to $1000 per genome (http://www.illumina.com). This remarkable development means affordable large-scale human genome sequencing projects are a reality, which in turn elevates the potential for improved understanding of variant-function relationships and the identification of causative variant/genes for genetic traits and diseases.

Even if new technologies can facilitate the completion of genome data with minimal cost, the main difficulty in genome studies is indeed the multi-filtration and iterative process for the identification of candidate genes and causal variants (**Figure 6.1**) [7]. Each person harbors about three million genomic variations; the variants are present throughout the genome [1]. Proper analyses can help narrow down millions of whole genome variants to less than ten SNPs which have strong clinical significance.

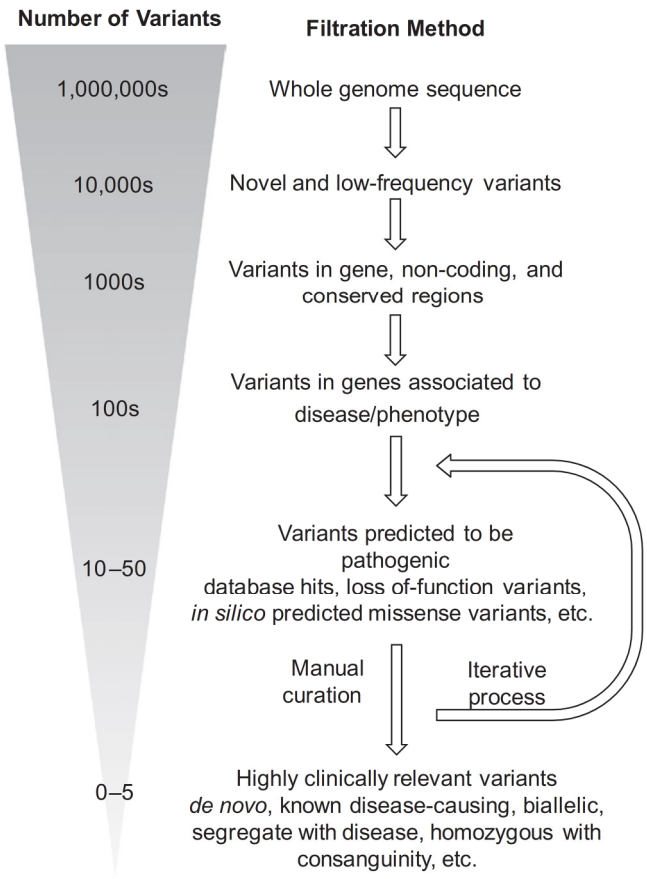


Figure 6.1: Variant analysis and filtration in whole genome sequencing. The identification of candidate genes and causal SNPs require multiple filtration steps (taken from [7]).

Although whole genome data can be informative, the functional annotation of variants in various genomic regions (coding, splice site, regulatory region, etc.) is quite complex—different types of functional variants have different weights. An example is illustrated by a complicated decision tree for analyzing candidate variants (**Figure 6.2**) [13]. This approach has been implemented in a bioinformatics tool, F-SNP [172], which uses conditional probabilities to assign the maximum likelihood that a SNP has significant functions at different genomic regions.

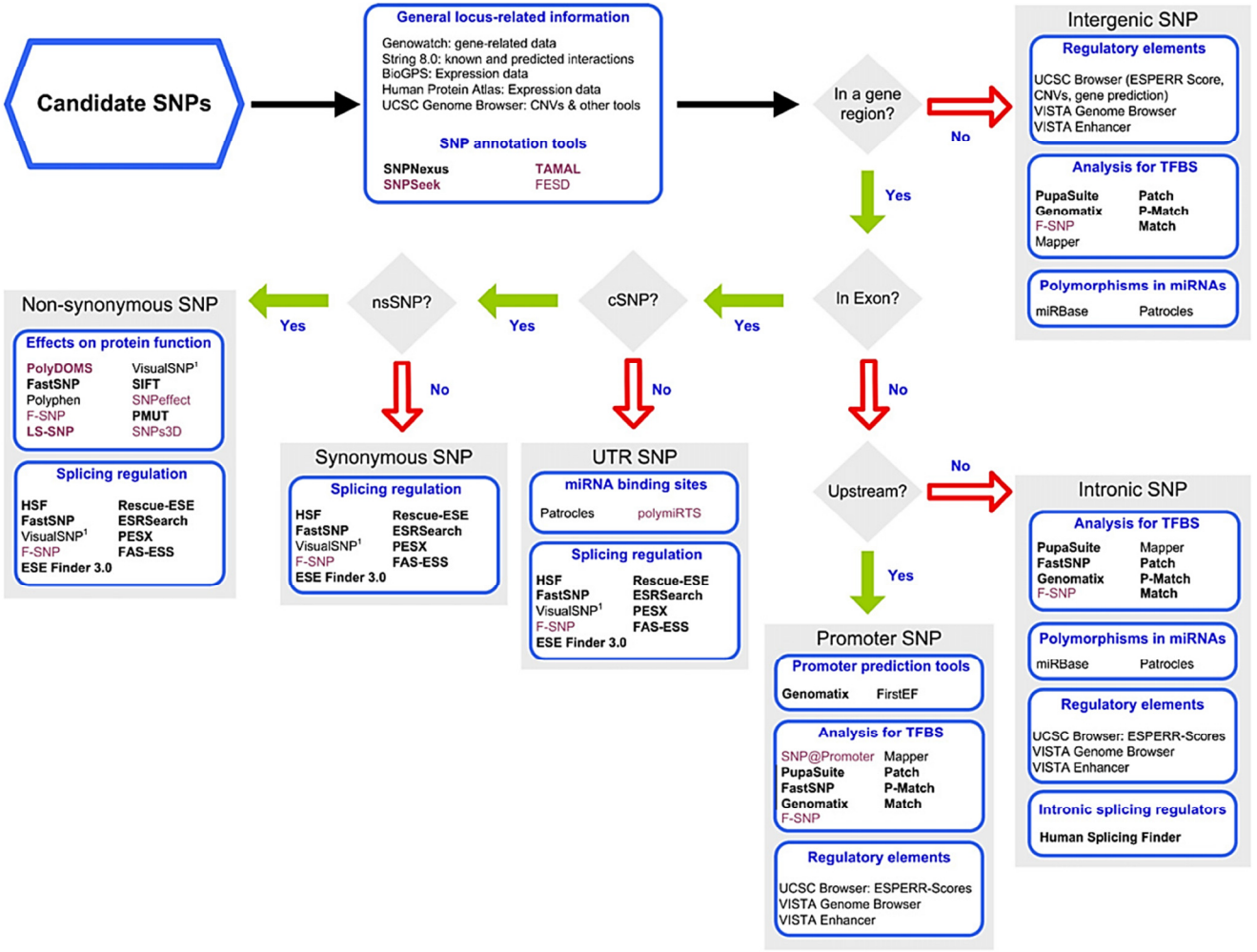


Figure 6.2: Tools for analyzing SNPs form different genome regions. Each box lists the common tools for analyzing certain type of genomic variants. The tools are largely redundant but somewhat different in the methodologies (taken from [13]).

Given that disease-causing variants are mostly located in the protein coding regions, the focus on analyzing coding variants can promote the detection of true causative factors. Above all, missense variants enable straightforward functional annotations since they underlie the best understood biological concept: DNA→RNA→protein.

Whole exome sequencing (WES) offers a cost effectiveness alternative approach to whole genome sequencing (WGS) for conducting the large-scale human genome research. While WGS reads an individual's entire DNA, exome sequencing targets only the protein coding regions (1% of the whole genome). More importantly, a whole exome is (for now) 1/6 the cost of whole genome and 1/15 the amount of data [173]. It is suggested that more than 10,000 exomes are needed to achieve statistical power [174]. Undoubtedly, the lower cost of WES can ensure more samples, better sequencing coverage, and better quality control of sequencing data. The relatively small amount of data facilitates exhaustive variant analysis and the possibility to combine effects of multiple variants together. As a result, the study of amino acid mutations caused by coding variants will greatly benefit from the advantages of WES.

Complexities and future directions for interpreting amino acid variants

Detailed knowledge about an individual variant is essential for the elucidation of molecular mechanisms that underlie their functional impact. The emerging practice in variant interpretation is to combine functional predictions from several sources to improve the annotation confidence and/or accuracy. At the moment, the accurate functional annotation of amino acid variants is still under development, since several key limitations can be foreseen: (1) confirmed functional data of coding variants are very

limited in terms of availability, quality and variability of experimental data. For proteins that bind to multiple ligands, it is also possible that mutagenesis tests only examine one substrate in the same *in vitro* experiment; the results present substrate-dependent effects as opposed to generalized functional significances of the variant. (2) The accurate prediction of protein-ligand interactions is limited by large computational resource requirements. (3) For genes that a have high number of genetic variants (such as genes in the CYP450 system), structural variations e.g. indels, copy number variants, and gene fusions which are coupled with missense mutations, introduce additional challenges for variant analysis [175].

Protein structural analysis may have a major impact in the next phase of genome studies as missense variants contribute to at least 60% of Mendelian disease development [5]. Computational prediction of various structural changes can be used to guide subsequent biological assays which will include functional implication of the specific amino acid mutations. Current trends in structural analysis use static protein data, e.g. stability change of the protein, physicochemical changes of the amino acid side chain, or changes in ligand binding environment to classify functional roles of a protein residue. However, it is very important not to ignore dynamic data since proteins are flexible biomolecules that can have functional interactions with substrates, biological compounds, or other proteins.

The strength of exhaustive molecular dynamic simulations for the ability to detect large-scale structural changes will increasingly become a desired method for *in silico* functional assessment of amino acid mutations at the level of tertiary and quaternary protein structures. For primary protein sequences, bioinformatics tools that combine the

conservation degree of amino acid substitutions, physicochemical/biological properties, and possibly structural-derived parameters are deemed essential for the preliminary process of variant prioritization.

Summary of this dissertation

Numerous genetic variations an individual carries joint with the potential impacts of some variants towards his/her health profiles create the demands for analyzing personal genome data. The ideal practice is that a variant assessment pipeline needs to be efficient and informative—using a systematic evaluation which is suitable for large genome data while providing a good combination of computational predictions (conservation-based and structural-based) and knowledge-based parameters in human genomics (disease and trait associations) [28]. In addition, a suitable user interface shall be implemented to promote the practical utility of the developing variant assessment protocol.

Although sequence conservation can be used to assess variant effects, it is not the only indicator for residue functionality. This dissertation asked if it is possible to overcome the weaknesses of the conservation-based assumption for variant deleteriousness. Specifically, can we combine the conservation-based approach with clinical information and structural data to attain a better understanding of the variant effects?

Using several data sources for the development of the three variant assessment pipelines (**Chapters 2, 4** and **5**), the dissertation highlights some of the difficulties yet to be overcome. The difficulties arise from several attributes: (1) limited number of curated gene-disease data. (2) The limitation of gene knowledge. (3) Diverse structural

characteristics of mutating proteins caused by disease causing mutations. (4) Errors in homology models. (5) Error in prediction programs.

The details for each observation are described below.

(1) *Limited number of curated gene-disease data.* The list of clinically curated gene-disease associations is not exhaustive, and the annotations often have different levels of significance (e.g. variants that were concluded from a single significant study, variants with moderate statistical associations, and variants derived from case-by-case reports, non-significant studies, or in vitro data). Notably, SNP databases frequently list associated diseases without specifying the annotation confidences.

(2) *The limitation of gene knowledge.* My research examined variant effects in three complicated situations. The analysis of personal genomes (**Chapter 2**) focused on homozygous missense variants (highly penetrant effects are most likely to be recessive) and identified 60-84 known homozygous nsSNPs per genome (across 40-77 genes). In addition, some individuals also carry *de novo* homozygous variants (range 1-5 variants per genome). Genomic variants of healthy individuals are less likely to have observable effects, although some variants are predicted to be deleterious. Without additional information of clinical data, the deleterious prediction is not as informative as expected. The AACDS classification scheme (**Chapter 2**) provides a concise description (and eight-level category) of a variant from the perspectives of consensus sequence-based deleterious prediction, types of mutation (disease-associated vs. neutral), and information on disease- or trait-associations with the gene.

The analysis of variants in epilepsy disorders suffered from very minimal information on the gene functions towards the disease development (**Chapter 4**). Literature searches indicated only 1 gene, among 68 genes of our dataset, may be linked to epilepsy. In addition, the number of candidate case variants is only one per a gene. The two limitations created difficulties in incorporating prior knowledge of gene biology to the functional assessments.

(3) *Diverse structural characteristics of mutating proteins caused by disease-causing mutations.* While ~75% of amino acid changes leading to Mendelian diseases (single gene diseases) consistently induce protein destabilization [37], the structural evaluation of missense variants in complex diseases (multiple gene diseases) is not that simple. The high contributions of missense variants in Mendelian genes on protein structures provide a better interpretation of variants effects than a list of the same variants that are linked to complex disorders—multiple variations in one gene and/or of several genes may work synergistically to create protein malfunction.

Furthermore, *in vivo* activities and/or structural characteristics of mutating proteins may be dissimilar between different types of disease-causing mutations. A study on protein kinase [176] indicates germline disease causing mutations mostly affect the substrate binding sites or protein-protein interaction sites, while cancer causing somatic mutations tend to altered ATP binding sites or catalytic residues. Bearing that some disease genes may harbor both types of causal mutations, but given the small number of causal variants per a gene, the sorting of causal variants into two distinct groups is not possible. The combination of

structural disturbances maybe diluted when multiple structural changes from various disease types were combined during the development of the prediction algorithm (**Chapter 5**).

(4) *Errors in homology models*. The understanding of genotype-phenotype relationships can be improved upon the examination of a mutated residue in its 3D protein context. Homology modeling is fulfilling the largely unavailable known protein structures. Homology models have with a wide range of accuracy—depending upon the sequence identity to the structural templates. Errors of modeled proteins exist (and are recognized), but they should not prevent the roles of computationally derived structures to assist functional annotations of amino acid variants. Errors from homology models include misplaced side chains, inaccurate loop modeling, distortions of protein core, or even wrong folds [177]. The model quality can be assured by performing structural minimizations (to resolve steric contacts) and assess the models using a series of quality scores (to evaluate the overall geometrical accuracy).

(5) *Error in prediction programs.* Many bioinformatics tools for predicting variant deleteriousness have been established, each with their own strengths. Nonetheless, the output of one tool may not be entirely reliable. During the analysis of variants in pharmacogenes (**Chapter 5**), large disagreements among six conservation-based indicators were discovered. More importantly, despite the fact that high impact variants in pharmacogenes contain extensive available clinical data, most prediction programs underestimated the damaging effects of pharamacogene variants (although they work well in other cases).

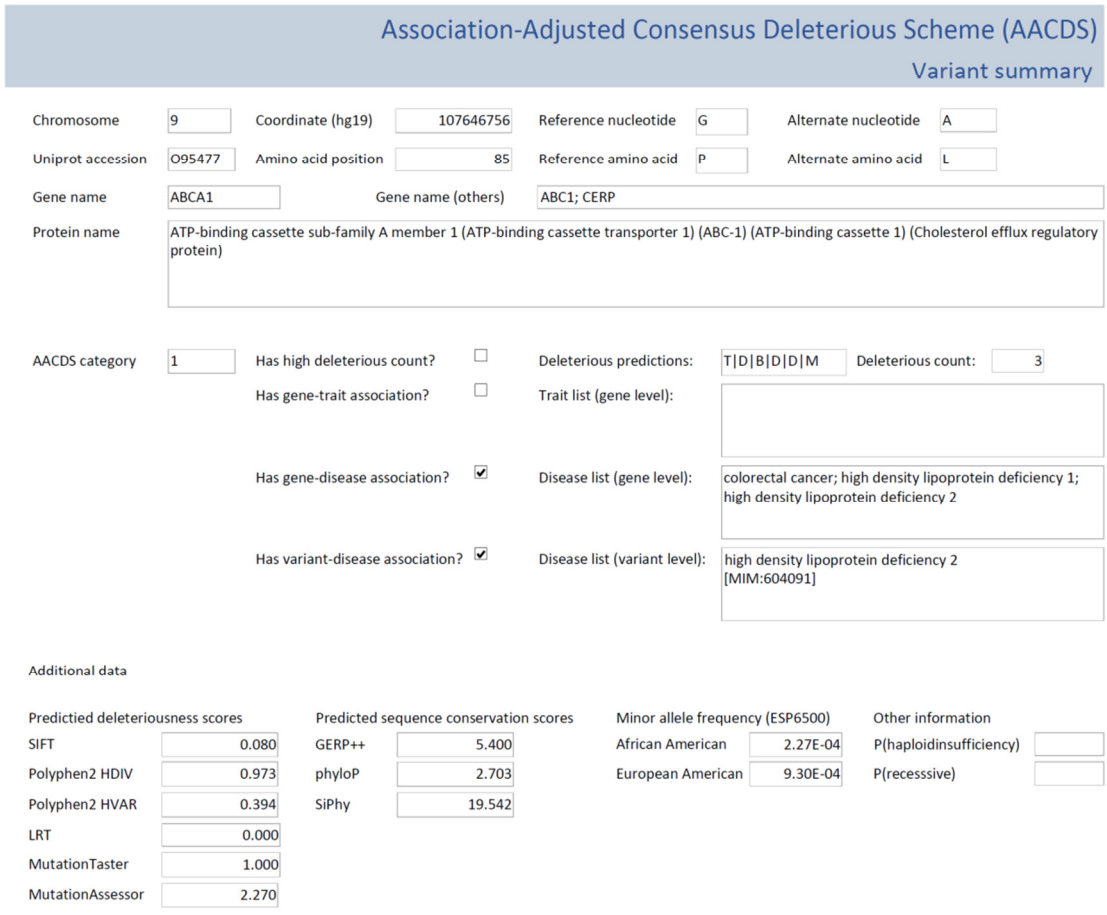
Despite these difficulties, this dissertation shows that the integrative approach for investigating variant-function relationships can be applied to at least three aspects of genome studies—personal genomics, genomics of epilepsy disorders, and genomics of variable drug responses. More importantly, the variant evaluation pipeline were implemented in a systematic manner, therefore, it is now possible to evaluate the large number of variants at a genome level (**Chapters 2**), in a disease-wise perspective (**Chapter 4**), and in a set of highly significant pharmacogenes (**Chapter 5**). Each implementation serves as a filter to identify functional significant variants which are worth clinical attentions (**Chapter 2**), which can guide subsequent experimental validations (**Chapter 4**), and which may induce the variability in drug responses (**Chapter 5**). The development of a database-driven web application for the AACDS classification scheme (**Chapter 3**) and the future development of a database for variants in pharmacogenes enhance the practical utility of research outcomes since minimal experience/expertise in genome interpretation is required.

The incorporation of knowledge of clinical associations and protein structure data serve as complementary tools to the existing conservation-based algorithms for variant deleteriousness. My efforts in generating a large number of high quality homology models and the employment of diverse structural analyses offer a way to detect common structural features which are induced/targeted by most functional variants (variants that lead to functional differences between wild type and mutant proteins). As a result, functional implications caused by any amino acid changes in a protein can be elucidated by a fast and systematic screening of structural disturbances. The approach provides an opportunity for investigating the joint effects of multiple structural changes from one or

many mutations in a specific protein and for establishing the likely causative variants in large genome data.

Overall, the dissertation explains essential developments in genome analysis. It is the first introduction of three variant assessment pipelines that utilize an efficient way to catalog missense variants on a large scale.

APPENDIX A: SUPPLEMENTARY INFORMATION FOR CHAPTER 2



Association-Adjusted Consensus Deleterious Scheme (AACDS)

Variant summary

Chromosome	9	Coordinate (hg19)	107646756	Reference nucleotide	G	Alternate nucleotide	A
Uniprot accession	O95477	Amino acid position	85	Reference amino acid	P	Alternate amino acid	L
Gene name	ABCA1	Gene name (others)	ABC1; CERP				
Protein name	ATP-binding cassette sub-family A member 1 (ATP-binding cassette transporter 1) (ABC-1) (ATP-binding cassette 1) (Cholesterol efflux regulatory protein)						

AACDS category	1	Has high deleterious count?	☐	Deleterious predictions:	T|D|B|D|D|M
		Has gene-trait association?	☐	Trait list (gene level):	
		Has gene-disease association?	☑	Disease list (gene level):	colorectal cancer; high density lipoprotein deficiency 1; high density lipoprotein deficiency 2
		Has variant-disease association?	☑	Disease list (variant level):	high density lipoprotein deficiency 2 [MIM:604091]

Additional data

Predictied deleteriousness scores		Predicted sequence conservation scores		Minor allele frequency (ESP6500)		Other information	
SIFT	0.080	GERP++	5.400	African American	2.27E-04	P(haploidinsufficiency)	
Polyphen2 HDIV	0.973	phyloP	2.703	European American	9.30E-04	P(recesssive)	
Polyphen2 HVAR	0.394	SiPhy	19.542				
LRT	0.000						
MutationTaster	1.000						
MutationAssessor	2.270						

Figure A.1: AACDS summary report. The report is provided to the user with the AACDS category of the variant and its relevant information, along with additional variant data.

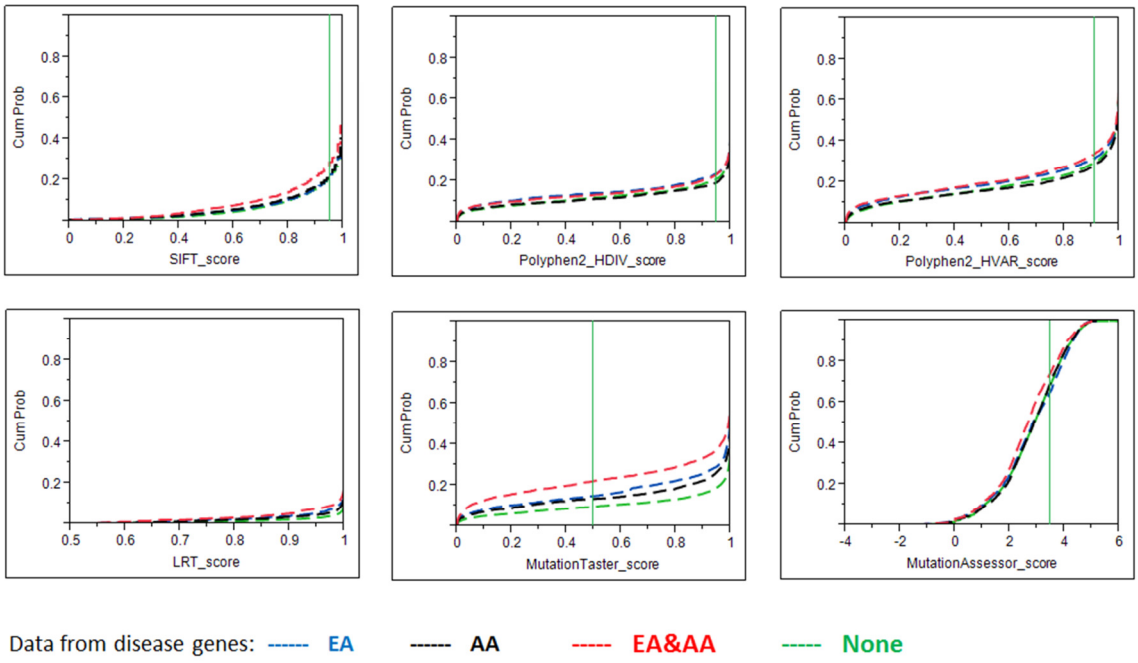


Figure A.2: Cumulative distribution plots for the six deleterious prediction scores. The X-axis represents the prediction scores, ordered by deleteriousness such that low and high scores for each prediction algorithm indicate neutral and damaging nsSNPs, respectively. For each prediction program, the score threshold for defining damaging SNPs is indicated by a vertical green line (threshold for LRT is at 0.999). The genes were classified into four groups depending upon population prevalence of their SNPs, using the difference in minor allele frequencies (MAFs) (cut-off of ± 5%) between European American (EA) and African American (AA) populations. The four gene groups are EA bias, AA bias, EA&AA bias, and no bias. For each plot, the dashed lines illustrate the cumulative distribution of deleterious prediction scores for disease-causing SNPs located in each gene group. The numbers of genes and SNPs are as follows: EA bias (222 genes, 3409 SNPs), AA bias (368 genes; 4,825 SNPS), EA&AA bias (234 genes; 4,225 SNPs), and no bias (965 genes; 12,214 SNPs). All disease-causing nsSNPs were retrieved from MSV3d [43] and SwissVar [44]. Population-specific minor allele frequencies for the variants were derived from NHLBI GO Exome Sequencing Project (ESP6500) (June 2012 release) [50].

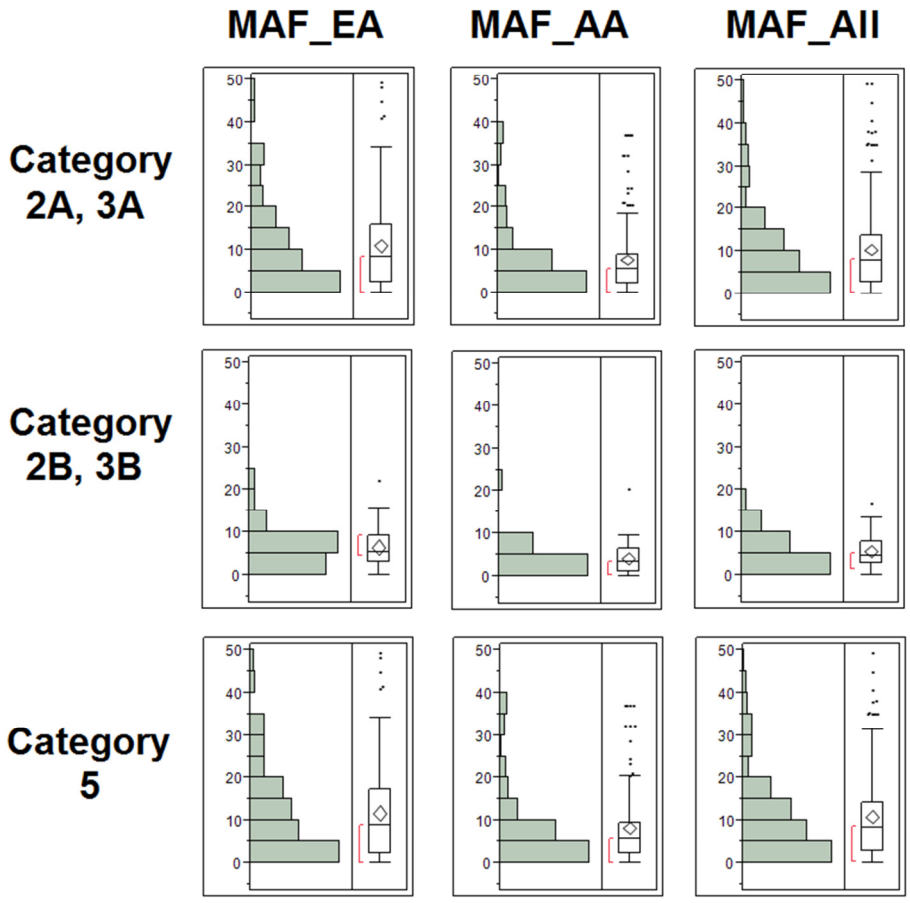


Figure A.3: Allele frequency distributions by AACDS score. The three columns indicate the minor allele frequency (MAF) in percent, listed in the order of European American (EA), African American (AA) and all populations (All). Only SNPs with available allele frequency data are represented here and the numbers in each group are 221, 33 and 165, respectively.

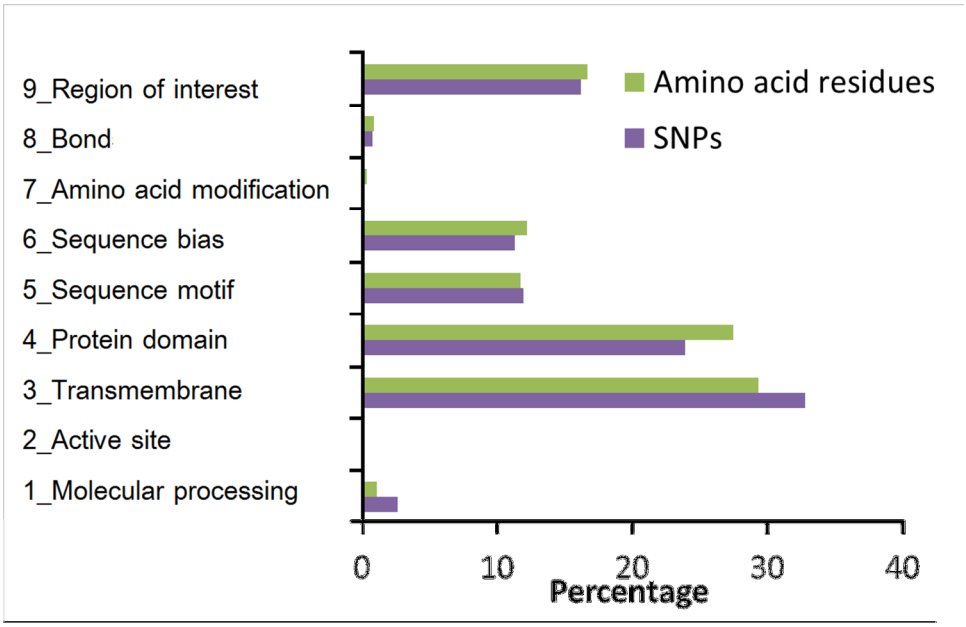


Figure A.4: Proportions of the 9 types of annotated protein regions found in all residues in the analyzed proteins vs. in SNP residues. Data were compiled from a set of 520 proteins whose sequence features are available from UniProt database [51].

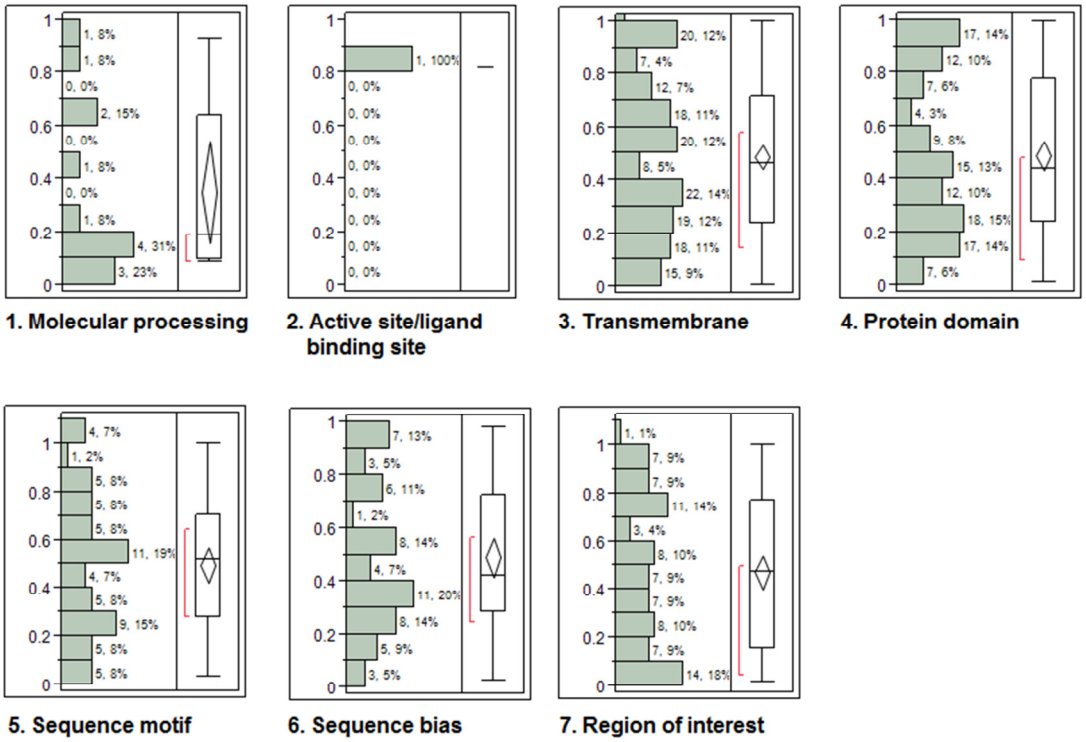


Figure A.5: Location of SNPs within proteins according to sequence feature type. A relative location near zero indicates the SNP is located near the N-terminus of that sequence feature. For clarity, a few features were excluded due to small sample sizes.

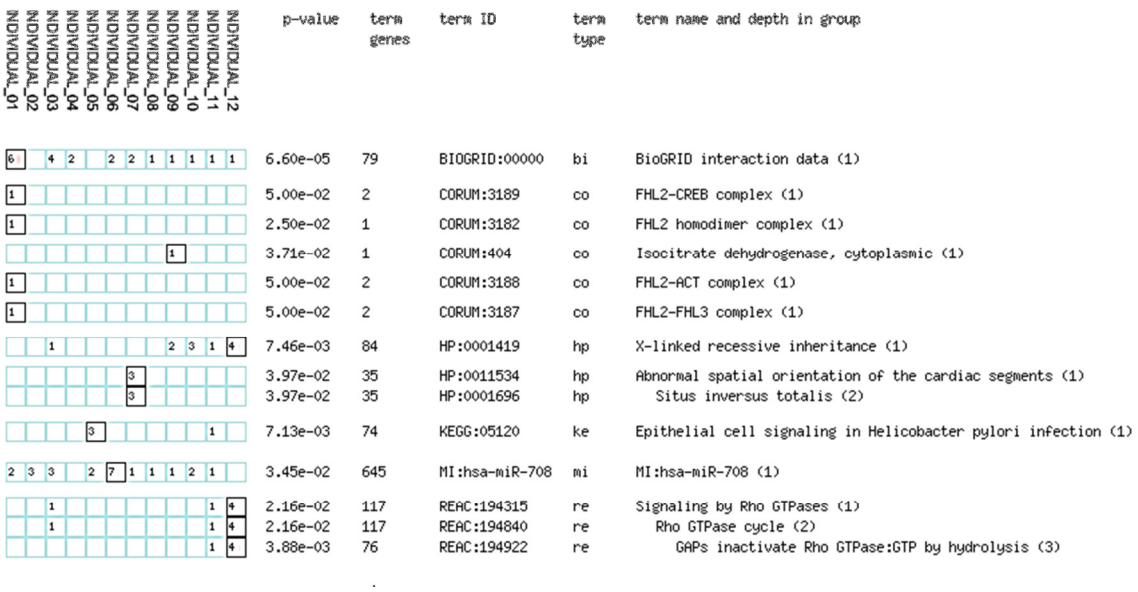


Figure A.6: Comparison of gene functional enrichment in the 12 genomes. The analysis was performed with g:Cocoa [78]. Each cell in the left most column indicates the number of queried genes from each individual that are associated with each annotation term. The highlighted cells indicate significant enrichment. The enrichment p-values are determined by the default multiple testing correction procedure g:SCS. The column "Term genes" indicates the total number of genes associated to each functional term. *Abbreviations: bi* BioGRID protein-protein interaction network, *co* CORUM protein complexes, *hp* human disease genes from Human Phenotype Ontology, *ke/re* KEGG/REACTOME pathway, *mi* MicroCosm microRNA sites.

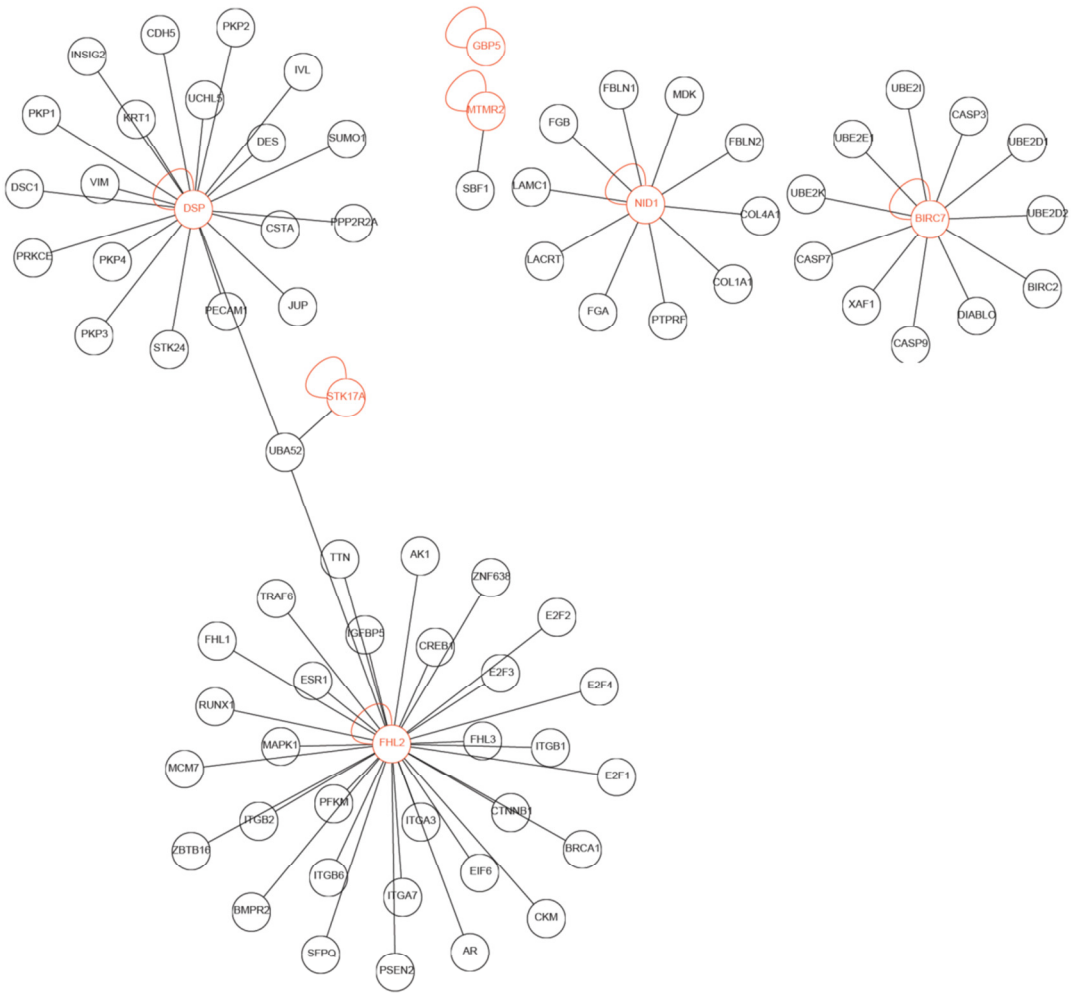


Figure A.7: BioGRID network for protein interactions in one person's genome. The red nodes highlight a subset of the query that is connected by an edge in the network. The black nodes are the immediate neighbors of the red nodes. The one private homozygous nsSNP from this individual is found in the *STK17A* gene.

Table A.1: List of protein structures used in supervised structural analysis.

Structure type (source)	Protein	Protein full length	Structure coverage		Details of PDB structures	Details of homology models		
			Range	%	PDB [resolution]	Template PDB [%seq ID]	QMEAN6 score (Z-score)	ModFOLD Score (P-value)
NMR (PDB)	APOE	317	19-317	94	2l7bA [n/a]	-	-	-
X-ray (PDB)	AMY2A	511	17-511	97	3oleA [1.55 Å]	-	-	-
X-ray (PDB)	CHIA	476	22-398	79	3fy1A [1.70 Å]	-	-	-
X-ray (PDB)	MTMR2	643	74-586	80	1lw3A [2.30 Å]	-	-	-
X-ray (PDB) & model (PMP)	THBS1	1170	549-1169	53	1ux6A [1.90 Å]	1yo8A [77%]	0.745 (-0.20)	0.3664 (2E-2)
Model (PMP)	MYH6	1939	3-777	38	-	4db1B [85%]	0.644 (-1.32)	0.6001 (1.61E-4)

Abbreviations: PDB RCSB Protein Data Bank [60], *PMP* Protein Model Portal [61], *Phyre* Protein Homology/analogY Recognition Engine [101].

Structure coverage represents the range of protein residues present in the 3D structure and the percentage of structural coverage with respect to the full length of the proteins. Each PDB accession number is indicated by its four letter code followed by the chain ID (in capital letter).

Details of homology models include the template' PDBs, sequence identity percentage between the target and the template sequences (in case of single-template modeling), or percentage of residues modeled at >90% confident (in case of multi-template modeling). Quality of homology models is evaluated by QMEAN6 [63] and ModFOLD4 [64] scores. QMEAN6 is a reliability score derived from a linear combination of six terms. It estimates model reliability between 0 and 1 (1 represents the best model). QMEAN Z-score for a given model is a comparable quality score to experimental structures of similar size. Models of low quality will have strongly negative QMEAN Z-scores. ModFOLD's global model quality score ranges between 0 and 1. In general, scores greater than 0.4 generally indicate more complete and confident models, which are highly similar to the native structure. ModFOLD's P-value represents the probability that each model is incorrect.

Table A.2: List of protein structures used in automated structural analysis.

Structure type (source)	Protein	Variant residue	Protein full length	Covered residues	% Coverage	PDB [resolution]
model (PDB)	OPSR	236	364	1-364	100	1kpxA [n/a]
x-ray (PDB)	1A03	121	365	25-298	75	3rl2A [2.39 Å]
x-ray (PDB)	1A11	91	365	25-299	75	1x7qA [1.45 Å]
x-ray (PDB)	ADA	8	363	5-363	99	3iarA [1.52 Å]
x-ray (PDB)	AL4A1	470	563	23-563	96	3v9hA [2.4 Å]
x-ray (PDB)	AMYP	145	511	17-511	97	3oliA [1.5 Å]
x-ray (PDB)	ARSB	376	533	42-533	92	1fsuA [2.5 Å]
x-ray (PDB)	C1S	119	688	16-174	23	1nziA [1.5 Å]
x-ray (PDB)	CCL8	49	99	25-99	76	1esrA [2.0 Å]
x-ray (PDB)	CP2A6	160	494	30-494	94	2fdvA [1.65 Å]
x-ray (PDB)	CSF1R	245	972	20-295	28	4dkdC [3.0 Å]
x-ray (PDB)	DRB5	149	266	30-219	71	1fv1B [1.9 Å]
x-ray (PDB)	FCG2A	63	317	37-207	54	1fcgA [2.0 Å]
x-ray (PDB)	HGFA	644	655	393-646	39	1yc0A [2.6 Å]
x-ray (PDB)	IDHC	178	414	4-410	98	3inmA [2.1 Å]
x-ray (PDB)	IF16	723	785	575-766	24	3rloA [1.8 Å]
x-ray (PDB)	KC1G2	173	415	44-338	71	2c47A [2.4 Å]
x-ray (PDB)	LYAM1	193	372	39-194	42	3cfwA [2.2 Å]
x-ray (PDB)	MGA	755	1857	93-954	46	3l4yA [1.8 Å]
x-ray (PDB)	MICA	174	383	24-297	72	1hyrC [2.7 Å]
x-ray (PDB)	PUR6	141	425	7-425	99	2h31A [2.8 Å]
x-ray (PDB)	PYGL	222	847	23-830	95	1l5rA [2.1 Å]
x-ray (PDB)	SNX7	59	387	25-150	33	3iq2A [1.7 Å]
x-ray (PDB)	SUN2	671	717	521-717	27	3unpA [2.39 Å]
x-ray (PDB)	THBG	303	415	40-414	90	2ceoA [2.8 Å]

Abbreviations: PDB RCSB Protein Data Bank [60].

Structure coverage represents the range of protein residues present in the 3D structure and the percentage of structural coverage with respect to the full length of the proteins. Each PDB accession number is indicated by its four letter code followed by the chain ID (in capital letter).

Table A.3: List of all private variants in the 12 genomes.

Subject ID	#SNPs (#genes)	Position, Base change (AA change)	Gene	Remarks	Gene-disease associations	Gene-trait associations	Protein sequence/structural features of variant	AACDS category
1	2 (2)	3:14106202 *G-->A (G177S)*	*TPRXL*	Putative protein		Ovarian reserve	Compositional bias (Ser-rich=76%)	3A, 5
		7:43622870 *G-->A (G10S)*	*ST17A*				N-terminal	6
2	3 (3)	17:43319304 *C-->T (P559L)*	*FMNL*				Interdomain linker	6
		17:72839464 *G-->C (P938A)*	*NMDE3*				Topological domain (cytoplasmic side)	6
		6:32009621 *C-->T (R4232Q)*	*TENX*		Tenascin-X deficiency	Phospholipid levels (plasma); HIV-1 control; Systemic lupus erythematosus	C-terminal	2B, 3B
3	3 (2)	1:148010896 *C-->G (V576L)*	*NBPFE*	Young gene, generated by gene duplications during primate evolution		AIDS progression	1 of the 10 NBPF domains	3A, 5
		1:148010901 *A-->C (L574W)*	*NBPFE*					3A, 5
		10:23384581 *C-->T (T15I)*	*MSRB2*				Transit peptide	6
4	1 (1)	1:111217425 *C-->T (E3K)*	*KCNA3*				N-terminal	6
5	2 (2)	11:4967679 *A-->T (S218T)*	*O51A4*				Transmembrane	4
		9:69424057 *C-->G (L785V)*	*A20A4*	Putative protein			Coiled-coil	6
6	none							
7	4 (4)	19:55998135 *G-->T (A145S)*	*NAT14*				N-acetyltransferase domain	6
		2:131976262 *A-->G (N96S)*	*POTEE*	Putative protein			N-terminal	6

Supplementary Table A.3 (continued)

Subject ID	#SNPs (#genes)	Position, Base change (AA change)	Gene	Remarks	Gene-disease associations	Gene-trait associations	Protein sequence/structural features of variant	AACDS category
7	4 (4)	2:43451823 *T-->C (T374A)*	*TISD*					6
		9:69423770 *C-->T (S689L)*	*A20A4*	Putative protein			Coiled-coil	6
8	none							
9	2 (2)	7:106300898 *G-->A (P149S)*	*CC71L*	Putative protein			Compositional bias (Pro-rich=34%)	6
		X:3229609 *G-->A (T2212M)*	*MXRA5*	Chromosome Y also has this pseudogene.			Ig-like C2-type 6 domain	4
10	2 (2)	4:9251554 *A-->T (D400V)*	*U17LI*					6
		9:69424057 *C-->G (L785V)*	*A20A4*	Putative protein			Coiled-coil	6
11	5 (5)	1:145333904 *A-->T (Q804L)*	*NBPFA*				NBPF 6 domain	6
		2:97911728 *G-->A (D1802N)*	*AN36A*	Putative protein				6
		4:9251554 *A-->T (D400V)*	*U17LI*					6
		X:154158785 *C-->G (E1094Q)*	*FA8*	The gene contains at least 53 genetic variants in families with hemophilia A.	Hemophilia A		Part of Factor VIIIa heavy chain	2A, 5
		X:70329183 *G-->A (R218C)*	*IL2RG*		Severe combined immunodeficiency X-linked T-cell-negative/B-cell-positive/NK-cell-negative; Agammaglobulinemia Swiss type; X-linked combined immunodeficiency		Topological domain (extracellular side), Fibronectin type-III domain	2A, 5

Supplementary Table A.3 (continued)

Subject ID	#SNPs (#genes)	Position, Base change (AA change)	Gene	Remarks	Gene-disease associations	Gene-trait associations	Protein sequence/structural features of variant	AACDS category
12	5 (5)	19:46394260 *C-->T (R274Q)*	*MYPOP*				Compositional bias (Pro-rich=31%)	4
		X:100169803 *C-->T (G292S)*	*XKR2*	Putative protein			Interhelix-linker	6
		X:34962803 *G-->A (D619N)*	*FA47B*					6
		X:37026758 *G-->A (S92N)*	*FA47C*	Putative protein				6
		X:38013836 *G-->T (Q364K)*	*SRPX*					6

The "Remarks" column describes the interesting features of gene annotations, obtained from UniProt database [51]. The SNPshot text-mining tool for PubMed abstracts [80] was used to explore if any of the private homozygous nsSNP-containing genes have clinical or experimental evidence for gene-drug or gene-disease associations.

Table A.4: List of all Categories 2A/2B variants affecting the same gene in more than one individual.

Gene	# SNPs	Associated-diseases	Associated-traits	Variant categories	Amino acid mutations	Note
I12R1	3	Mendelian susceptibility to mycobacterial disease		2A	R156H	same variant
FRMD7	3	Nystagmus congenital x-linked type 1		2A	R468H	same variant
FRAS1	3	Fraser syndrome	Hair morphology	2A, 3A | 2A, 3A | 2A, 3A	T2203I | G2230R | A2251T	
MUC5B	3	Pulmonary fibrosis idiopathic		2A | 2A | 2A	R2211P | P2830L | S3061P	
ZC12D	2	Sporadic lung cancer		2A | 2A	E51K | P405S	
CO6A5	2	Atopic dermatitis		2A | 2A	V1276I | T1280P	
CUBN	2	Breast cancer; Colorectal cancer; Renal cell carcinoma case; Recessive hereditary megaloblastic anemia 1	Folate pathway vitamin levels; MRI atrophy measures	2A, 3A | 2A, 3A	I2984V | E3002G	
FA9	2	Recessive x-linked hemophilia b; Thrombophilia		2A | 2A	T194A	same variant
ADA	2	Pancreatic ductal adenocarcinoma; Severe combined immunodeficiency autosomal recessive		2A | 2A	D8N | K80R	
OPSR	2	Partial colorblindness protan series; Blue cone monochromacy		2A | 2A	A174V | M236V	
MYH6	2	Atrial septal defect type 3; Familial hypertrophic cardiomyopathy type 14; Cardiomyopathy dilated type 1ee; Sick sinus syndrome type 3	Resting heart rate; Electrocardiographic trait	2B, 3B | 2A, 3A	G56R | A1130T	
PGCA	2	Spondyloepiphyseal dysplasia type kimberley; Spondyloepimetaphyseal dysplasia aggrecan type; Osteochondritis dissecans short stature and early-onset osteoarthritis	Height	2A, 3A | 2A, 3A	D1142E | E1294D	
LAMB3	2	Colorectal cancer; Epidermolysis bullosa junctional herlitz type; Generalized atrophic benign epidermolysis bullosa		2A	M852L	same variant
MICA	2	Progression of monoclonal gammopathy of undetermined significanceto multiple myeloma; Psoriasis type 1; Psoriatic arthritis	Rheumatoid arthritis; HIV-1 control; Hepatocellular carcinoma	2A, 3A | 2A, 3A	M174V | P294A	
C1GLC	2	TN syndrome		2A | 2A	D131E | A143V	

Supplementary Table A.4 (continued)

Gene	# SNPs	Associated-diseases	Associated-traits	Variant categories	Amino acid mutations	Note
GCP6	2	Microcephaly with chorioretinopathy		2A | 2A	R1763W | A884V	
SP110	2	Breast cancer; Hepatic venoocclusive disease with immunodeficiency		2A | 2A	E207K | A128V	
VPP4	2	Distal renal tubular acidosis with preserved hearing	F-cell distribution	1, 2A, 3A	M580T	same variant

#SNPs represents the total number of homozygous nsSNP found in each gene from all 12 individuals. Multiple diseases or traits that are associated with the genes are separated with ";". The variant category and the amino acid mutation for each nsSNP are partitioned with "|".

Table A.5: List of all Category 2B variants.

Gene	Position	Base change (AA change)	Del count	Con count	Grantham score	Protein stability change	Site annotations	Associated disease(s)
GNAT2	1:110151395	*G-->T (L107I)*	3	3	5	Decrease		Achromatopsia type 4
IDH1	2:209108317	*C-->T (V178I)*	4	3	29	Neutral		Glioma
PRSS12	4:119203221	*C-->T (R833Q)*	3	3	43	Neutral	DOMAIN	Mental retardation autosomal recessive type 1
LAMA2	6:129824406	*A-->G (N2843S)*	4	3	46	Decrease	DOMAIN	Merosin-deficient congenital muscular dystrophy type 1A
SYNE1	6:152443744	*G-->T (L8741M)*	3	3	15	Neutral	TOPO_DOM; DOMAIN	Spinocerebellar ataxia autosomal recessive type 8; Autosomal recessive cerebellar ataxia type 1; Emery-Dreifuss muscular dystrophy type 4
DNAH11	7:21778429	*T-->C (Y2593H)*	5	3	83	Neutral	REGION (AAA 3 (By similarity))	Kartagener syndrome; Primary ciliary dyskinesia type 7
RP9	7:33134883	*T-->C (K210R)*	3	3	26	Neutral	COMPBIAS	Retinitis pigmentosa type 9
ELN	7:73474825	*G-->C (G610R)*	3	3	125	Neutral	COMPBIAS	Cutis laxa, autosomal dominant, type 1; Supravalvular aortic stenosis
SPTLC1	9:94830356	*C-->A (R151L)*	3	3	102	Neutral	TOPO_DOM	Hereditary sensory and autonomic neuropathy type 1A
MYH6	14:23876267	*C-->T (G56R)*	4	3	125	Neutral	DOMAIN	Atrial septal defect type 3; Familial hypertrophic cardiomyopathy type 14; Cardiomyopathy dilated type 1EE; Sick sinus syndrome type 3
APOE	19:45412079	*C-->T (R176C)*	5	3	180	Decrease	REPEAT	Hyperlipoproteinemia type 3; Familial dysbetalipoproteinemia; Sea-blue histiocyte disease; Lipoprotein glomerulopathy
NLRP12	19:54313707	*G-->C (F402L)*	3	1	22	Neutral	DOMAIN	Familial cold autoinflammatory syndrome type 2
TRMU	22:46731689	*G-->T (A10S)*	5	3	99	Decrease	NP_BIND	Transient infantile liver failure
TBX22	X:79281202	*G-->A (E187K)*	5	3	54	Neutral	DNA_BIND	X-linked cleft palate with ankyloglossia

Table A.6: List of all Category 3B variants.

Gene	Position	Base change (AA change)	Del count	Con count	Grantham score	Protein stability change	Site annotations	Associated trait(s)
NVL	1:224482084	*C-->T (V404I)*	4	3	29	Neutral		Major depressive disorder
SRBD1	2:45640334	*T-->C (K811R)*	4	3	26	Neutral		Glaucoma
ZNF385D	3:21706469	*G-->T (P25H)*	3	3	77	Neutral		Partial epilepsies
DNAH1	3:52390789	*C-->T (R1285W)*	3	3	101	Decrease	REGION (Stem (By similarity))	Bipolar disorder
DNAH1	3:52429665	*C-->T (R3809C)*	3	3	180	Decrease	REGION (AAA 6 (By similarity))	Bipolar disorder
HLA-DRB5	6:32489949	*G-->A (R35C)*	3	3	180	Increase	TOPO_DOM; REGION (Beta-1)	Chronic lymphocytic leukemia; Ulcerative colitis; Parkinson's disease
LAMA2	6:129824406	*A-->G (N2843S)*	4	3	46	Decrease	DOMAIN	Body mass index
SYNE1	6:152443744	*G-->T (L8741M)*	3	3	15	Neutral	TOPO_DOM; DOMAIN	Bipolar disorder and major depressive disorder (combined); Bipolar disorder; Tonometry
LPA	6:160977059	*A-->C (N4165K)*	3	2	94	Decrease	DOMAIN	Protein quantitative trait loci; LDL cholesterol; HDL cholesterol; Total cholesterol; Lp (a) level; Coronary heart disease; Response to statin therapy (LDL-C)

Supplementary Table A.6 (continued)

Gene	Position	Base change (AA change)	Del count	Con count	Grantham score	Protein stability change	Site annotations	Associated trait(s)
DNAH11	7:21778429	*T-->C (Y2593H)*	5	3	83	Neutral	REGION (AAA 3 (By similarity))	Multiple myeloma; LDL cholesterol; Total ventricular volume; MRI atrophy measures; Total cholesterol
FRMD4A	10:13699338	*T-->G (T751P)*	5	3	38	Decrease	COMPBIAS	RR interval (heart rate); Alzheimer's disease
MRC1	10:18138519	*C-->G (P359A)*	3	3	27	Decrease	TOPO_DOM	Cardiovascular disease risk factors
ANKRD30A	10:37451768*	*T-->G (L665W)*	3	1	61	Decrease		Metabolite levels; Hemostatic factors and hematological phenotypes
PCDH9	13:67800867	*C-->T (S569N)*	3	3	46	Neutral	TOPO_DOM; DOMAIN	Obesity
MYH6	14:23876267	*C-->T (G56R)*	4	3	125	Neutral	DOMAIN	Resting heart rate; Electrocardiographic traits
HERC1	15:63988400	*G-->C (L1682V)*	4	3	32	Neutral		Iris characteristics
APOE	19:45412079	*C-->T (R176C)*	5	3	180	Decrease	REPEAT	Cardiovascular disease risk factor; Alzheimer's disease; Quantitative traits; Response to statin therapy (LDL-C); C-reactive protein
GGTLC1	20:23965995	*A-->T (V179E)*	4	3	121	Decrease		Erectile dysfunction and prostate cancer treatment
RTDR1	22:23482483	*G-->A (T42M)*	3	3	81	Neutral		Height
SYTL5	X:37935825	*T-->C (L187P)*	3	3	98	Neutral		Erectile dysfunction and prostate cancer treatment

* One SNP (10:37451768 T-->G) was observed in two individuals.

Table A.7: List of all Category 4 Variants.

Gene	Position	Base change (AA change)	Del count	Con count	Grantham score	Protein stability change	Site annotations
NBPF3	1:21809667	*C-->G (P564A)*	3	0	27	Neutral	DOMAIN
AMY2A	1:104161650	*C-->T (P145S)*	4	3	74	Decrease	
CHIA	1:111854859	*C-->T (R35W)*	3	2	101	Decrease	
ITGA10	1:145535814	*C-->T (R668W)*	4	3	101	Neutral	TOPO_DOM
QSOX1	1:180148012	*G-->C (G200A)*	3	3	60	Neutral	
SNTG2	2:1168781	*C-->A (S168Y)*	5	3	144	Increase	
CCDC142	2:74702756	*C-->T (R534Q)*	4	3	43	Neutral	
FOXD4L1	2:114257319	*C-->G (N162K)*	4	3	94	Neutral	DNA_BIND
UNC80	2:210798699	*G-->C (V1984L)*	3	3	32	Neutral	
SLC26A6	3:48669447	*C-->T (V206M)*	3	3	21	Neutral	TRANSMEM
DNASE1L3	3:58183636	*G-->A (R206C)*	5	3	180	Decrease	
KBTBD8	3:67053926	*T-->C (F179L)*	3	3	22	Neutral	DOMAIN
TKTL2	4:164393835	*A-->G (F351S)*	5	3	155	Decrease	
ENPP5	6:46135884	*C-->G (R39P)*	6	3	103	Neutral	
TCP10L2	6:167587284	*G-->C (R63T)*	3	1	71	Decrease	
RSPH10B	7:5983568	*A-->C (I528S)*	4	3	142	Decrease	
USP17L2	8:11995540	*C-->T (E244K)*	3	3	54	Decrease	
FER1L6	8:125115420	*G-->A (R1720Q)*	3	3	43	Neutral	TOPO_DOM
EPPK1	8:144944225	*C-->T (R1066H)*	3	3	29	Decrease	REPEAT
OR1N2	9:125316028	*T-->C (F194L)*	5	3	22	Neutral	TOPO_DOM
STAMBPL1	10:90673047	*G-->A (E204K)*	3	3	54	Neutral	
IFIT1B	10:91143374	*G-->T (A102S)*	4	3	99	Decrease	REPEAT
HABP2	10:115348046	*G-->A (G534E)*	5	3	98	Decrease	DOMAIN
NRAP	10:115410234	*T-->C (Y249C)*	4	3	194	Neutral	REPEAT
NPS	10:129350856	*G-->C (V75L)*	3	3	32	Neutral	
OR51A2	11:4976544	*A-->T (Y134N)*	4	3	143	Neutral	TOPO_DOM
OR52D1	11:5510598	*A-->T (Y221F)*	5	3	22	Neutral	TOPO_DOM
OR2D3	11:6942726	*G-->C (W165S)*	5	3	177	Neutral	TRANSMEM
ALDH3B2	11:67431914	*G-->A (R276W)*	3	0	101	Increase	
SCYL2	12:100708367	*C-->T (P357L)*	4	3	98	Neutral	
TEP1	14:20846927	*C-->G (G1780R)*	3	3	125	Neutral	REPEAT
THBS1	15:39882178	*A-->G (N700S)*	3	3	46	Neutral	REPEAT
SPESP1	15:69238272	*G-->T (L133F)*	3	3	22	Neutral	
ABCC11	16:48204078	*T-->A (N1277Y)*	5	3	143	Neutral	TOPO_DOM; DOMAIN

Supplementary Table A.7 (continued)

Gene	Position	Base change (AA change)	Del count	Con count	Grantham score	Protein stability change	Site annotations
PHLPP2	16:71683718	*A-->G (L1016S)*	4	3	145	Neutral	DOMAIN
PMFBP1	16:72184566	*C-->T (E193K)*	3	3	54	Neutral	
CTC1	17:8141897	*C-->G (S83T)*	4	3	58	Neutral	
KRT23	17:39092741	*C-->A (G39W)*	5	3	184	Neutral	REGION (Head)
KRT40	17:39135089	*G-->A (T388M)*	3	3	81	Increase	REGION (Rod)
KRTAP2-2	17:39211434	*G-->C (F10L)*	3	3	22	Decrease	REPEAT
KRTAP2-3	17:39216085	*C-->T (C73Y)*	4	3	194	Neutral	REPEAT
BTBD17	17:72353745	*G-->T (A163E)*	3	3	107	Neutral	
ENGASE	17:77078069	*G-->A (R321H)*	3	3	29	Neutral	DOMAIN
TBCD	17:80710097	*G-->T (G10C)*	3	3	159	Neutral	
PCSK4	19:1487195	*G-->A (T267M)*	3	3	81	Decrease	REGION (Catalytic, by similarity)
CSNK1G2	19:1978927	*G-->T (V173L)*	3	3	32	Decrease	DOMAIN
OR10H5	19:15905468	*T-->G (C204G)*	3	3	159	Neutral	TRANSMEM
OR10H5	19:15905505	*T-->C (L216P)*	3	3	98	Neutral	TRANSMEM
NXNL1	19:17566634	*T-->A (E154V)*	3	3	121	Neutral	DOMAIN
ZNF737	19:20727903	*T-->C (Y369C)*	3	1	194	Increase	ZN_FING
PSG1	19:43373078	*C-->G (W273S)*	3	3	177	Decrease	DOMAIN
ZNF534	19:52942535	*T-->G (C621G)*	4	3	159	Neutral	ZN_FING
SDCBP2	20:1292989	*C-->T (G242R)*	5	3	125	Decrease	DOMAIN
LRRN4	20:6033004	*G-->A (L148F)*	6	3	22	Neutral	TOPO_DOM; REPEAT
OTOR	20:16729138	*T-->C (L31P)*	4	3	98	Decrease	
DDX27	20:47859217	*G-->A (G766S)*	3	3	56	Neutral	
UMODL1	21:43504286	*G-->A (D138N)*	4	3	23	Neutral	TOPO_DOM; DOMAIN
HDX	X:83723541	*A-->G (F397S)*	3	3	155	Neutral	

Table A.8: Summary of automated structural analysis.

Source (# analyses)	Predictions	total # SNPs	# SNPs in each category*						
			2A	2B	3A	3B	4	5	6
SDM (n=33)	highly stabilizing/destabilizing ($\leq$ -2 or $\geq$ 2 kcal mol^{-1})	5	0	0	1	0	1	1	2
	stabilizing/destabilizing (between - 2 to -1 or 1 to 2 kcal mol^{-1})	5	2	0	2	1	0	0	0
	Slightly stabilizing/destabilizing (between -1 to -0.5 or 0.5 to 1 kcal mol^{-1})	8	2	0	0	0	1	2	3
	neutral (between ± 0.5 kcal mol-1)	15	4	1	9	0	0	0	1
Crystalographic B-factor of Cα atom (n=31)	small (B-factor <60)	27	6	1	10	1	2	2	5
	medium (B-factor ≥60 but <100)	4	0	0	2	0	0	1	1
	large (B-factor ≥100)	0	0	0	0	0	0	0	0
PredyFlexy (n=32)	rigid	19	5	0	5	0	0	3	6
	intermediate	9	1	1	5	0	2	0	0
	flexible	4	1	0	2	1	0	0	0
FlexPred (n=33)	conformationally flexible	3	0	0	1	1	0	0	1
	conformationally rigid	30	8	1	11	0	2	3	5
PDBe (n =25)	is a binding site	1	1	0	0	0	0	0	0
	is not a binding site	24	5	1	8	1	2	1	6
3DLigandSite (n=7)	is a binding site	0	0	0	0	0	0	0	0
PatchFinder (n=33)	Is part of conserved residue cluster	3	1	0	1	0	0	1	0
	Is not part of conserved residue cluster	30	7	1	11	1	2	2	6

The "source" column lists the name of prediction program or sources of structural information.

# of analyses refers to the number of 3D structures that were used to obtain the results. Some structures cannot be assessed by a certain approach, hence they are omitted.

Table A.9: List of X-linked recessive mutations.

Subject ID	Gene	Position	Base change (AA change)	rsID (%MAF EA/AA/All)	Disease/traits	Del count	Con count
3	*ARSE*	X:2856155	*C-->T (G424S)*	rs35143646 (32.9/17.9/49.2)	Chondrodysplasia punctata X-linked recessive type 1; Height	1	3
9	*ZNF674*	X:46360317	*T-->C (K236R)*	rs201621696 (0.8/0.1/0.6)	Mental retardation X-linked type 92	0	3
	BMP15	X:50658966	*G-->A (A180T)*	rs104894767 (1.4/0.3/1.0)	Ovarian dysgenesis type 2; X-linked hypergonadotropic ovarian dysgenesis or hypergonadotropic ovarian failure due to ovarian dysgenesis; Premature ovarian failure type 4	0	1
	C1GALT1C1	X:119760594	*G-->A (A143V)*	rs45557031 (2.5/0.5/1/8)	Tn syndrome	2	3
	FRMD7	X:131212642	*C-->T (R468H)*	rs6637934 (5.7/7.5/6.4)	Nystagmus congenital X- linked type 1	1	3
	F9	X:138633280	*A-->G (T194A)*	rs6048 (29.8/12.3/23.5)	Recessive X-linked hemophilia B;Christmas disease; Thrombophilia due to factor IX defect	0	2
	SYTL5	X:37985895	*G-->A (R702H)*	rs143176819 (0.03/0.0/0.02)	Erectile dysfunction and prostate cancer treatment	1	3
	TIMP1	X:47444361	*C-->T (P50S)*	rs145349279 (0.07/0.1/0.09)	Other erythrocyte phenotypes	0	1
10	*SERPINA7*	X:105278361	*C-->A (L303F)*	rs1804495 (11.2/12.3/11.6)	Thyroxine-binding globulin deficiency	2	2
	C1GALT1C1	X:119760629	*A-->T (D131E)*	rs17261572 (-/-/13.2)	Tn syndrome	0	2
	FRMD7	X:131212642	*C-->T (R468H)*	rs6637934 (5.7/7.5/6.4)	nystagmus congenital X- linked type 1	1	3
	F9	X:138633280	*A-->G (T194A)*	rs6048 (29.8/12.3/23.5)	Recessive X-linked hemophilia B; Christmas disease; Thrombophilia due to factor IX defect	0	2
11	*TBX22*	X:79281202	*G-->A (E187K)*	rs34244923 (6.8/0.8/4.6)	X-linked cleft palate with ankyloglossia	5	3
	CHM	X:85233820	*T-->A (S89C)*	rs145707160 (2.0/0.2/1.3)	Choroideremia	1	3
	ZCCHC16	X:111698440	*G-->T (D162Y)*	rs7474140 (21.9/6.8/16.4)	Biochemical measures	0	0
	OPN1LW	X:153420176	*A-->G (M236V)*	rs78093025 (9.5/36.7/19.4)	Partial colorblindness protan series;protanopia; Blue cone monochromacy	0	1

Supplementary Table A.9 (continued)

Subject ID	Gene	Position	Base change (AA change)	rsID (%MAF EA/AA/All)	Disease/traits	Del count	Con count
	SYTL5	X:37935825	*T-->C (L187P)*	rs144659697 (0.04/0.0/0.03)	Erectile dysfunction and prostate cancer treatment	3	3
	OPHN1	X:67652748	*C-->T (V39I)*	rs41303733 (7.8/1.5/5.5)	Mental retardation X- linked OPHN1-related	2	3
12	*OCRL*	X:128674722	*C-->T (T14I)*	rs61752970 (0.5/0.1/0.4)	Lowe oculocerebrorenal syndrome; Dent disease type 2	1	3
	FRMD7	X:131212642	*C-->T (R468H)*	rs6637934 (5.7/7.5/6.4)	Nystagmus congenital X- linked type 1	1	3
	OPN1LW	X:153418524	*C-->T (A174V)*	rs149897670 (7.6 /20.3/12.4)	Partial colorblindness protan series;protanopia; Blue cone monochromacy	0	1

APPENDIX B: SUPPLEMENTARY INFORMATION FOR CHAPTER 5

Table B.1: List of molecular functions and the numbers of drug partners for the 48 VIPs. Functional terms were analyzed with the g:Profiler web server (accessed December 17, 2013) [78] to summarize aspects of gene function.

Gene symbol	Protein name	Molecular function terms							# of drug pairs			
		Drug binding	Heme binding	Iron Ion binding	Oxygen binding	Beta-adrenergic receptor activity	Oxidoreductase activity	Electron carrier activity	Drug	Drug substrate	Drug Inhibitor	Drug Inducer
ABCB1	Multidrug resistance protein 1								77	68	40	
ACE	Angiotensin-converting enzyme	x							6			
ADH1A	Alcohol dehydrogenase 1A						x		1			
ADH1B	Alcohol dehydrogenase 1B						x		1			
ADH1C	Alcohol dehydrogenase 1C								1			
ADRB1	Beta-1 adrenergic receptor	x				x			15			
ADRB2	Beta-2 adrenergic receptor	x				x			2			
AHR	Aryl hydrocarbon receptor											
ALDH1A1	Retinal dehydrogenase 1						x					
ALOX5	Arachidonate 5-lipoxygenase			x			x		1			
BRCA1	Breast cancer type 1 susceptibility protein								1			
COMT	Catechol O-methyltransferase								2			
CYP1A2	Cytochrome P450 1A2		x	x			x	x		4	1	4
CYP2A6	Cytochrome P450 2A6		x	x			x	x		11		
CYP2B6	Cytochrome P450 2B6		x	x			x	x		18	4	7
CYP2C19	Cytochrome P450 2C19		x	x	x		x	x		25	14	4
CYP2C8	Cytochrome P450 2C8		x	x			x	x	9			
CYP2C9	Cytochrome P450 2C9	x	x	x			x	x	16			
CYP2D6	Cytochrome P450 2D6	x	x	x			x	x	38			
CYP2E1	Cytochrome P450 2E1		x	x	x		x	x	2			

Supplementary Table B.1 (continued)

Gene symbol	Protein name	Molecular function terms							# of drug pairs			
		Drug binding	Heme binding	Iron Ion binding	Oxygen binding	Beta-adrenergic receptor activity	Oxidoreductase activity	Electron carrier activity	Drug	Drug substrate	Drug Inhibitor	Drug Inducer
CYP2J2	Cytochrome P450 2J2		x	x			x	x	3			
CYP3A4	Cytochrome P450 3A4		x	x	x		x	x	72			
CYP3A5	Cytochrome P450 3A5		x	x	x		x	x	67	10		
DPYD	Dihydropyrimidine dehydrogenase						x		3			
DRD2	D(2) dopamine receptor	x							1			
F5	Coagulation factor V								4			
G6PD	Glucose-6-phosphate 1-dehydrogenase						x		11			
GSTP1	Glutathione S-transferase P								6			
GSTT1	Glutathione S-transferase theta-1						x			8		
HMGCR	3-hydroxy-3-methylglutaryl-coenzyme A reductase						x		5			
KCNH2	Potassium voltage-gated channel subfamily H member 2								43			
KCNJ11	ATP-sensitive inward rectifier potassium channel 11								5			
MTHFR	Methylenetetrahydrofolate reductase						x		6			
NQO1	NAD(P)H dehydrogenase						x		4			
NR1I2	Nuclear receptor subfamily 1 group I member 2	x							82			
P2RY1	P2Y purinoceptor 1											
P2RY12	P2Y purinoceptor 12								2			
PTGIS	Prostacyclin synthase		x	x			x	x				
PTGS2	Prostaglandin G/H synthase 2		x				x		6			
SCN5A	Sodium channel protein type 5 subunit alpha								11			
Total		7	13	13	2	24	4	12	546	147	62	15

Table B.2: A list of selected protein 3D structures, their data sources and the quality parameters.

Protein	Protein full length	Structure coverage†		Experimental structure	Homology model		
		Residue range	% of sequence coverage	PDB ID [method]	Template PDB [% confident]	QMEAN6 score (Z-score)	ModFOLD4 score (p-value)
ACE	1306	30-639, 645-1223	44.3	3nxqA [x-ray, 1.99 Å], 1uzeA [x-ray, 1.82 Å]			
ADH1A	375	1-375	100		multiple [100%]	0.856 (0.98)	0.725 (5.66E-05)
ADH1B	375	2-375	99.7	1u3uA [x-ray, 1.6 Å]			
ADH1C	375	2-375	99.7	1u3wA [x-ray, 1.45 Å]			
ADRB1	477	1-477	100		multiple [87%]	0.356 (-4.81)	0.351 (2.58E-03)
ADRB2	413	1-413	100		multiple [92%]	0.433 (-3.90)	0.417 (1.09E-03)
AHR	848	113-426	37.0		3gdiB [15.6%]	0.418 (-4.06)	0.237 (1.55E-02)
ALDH1A1	501	1-501	100		multiple [99%]	0.788 (0.25)	0.805 (3.14E-05)
ALOX5	674	1-674	100		multiple [100%]	0.686 (-0.85)	0.492 (4.69E-04)
BRCA1	1863	1-103, 1649-1859	16.9	1jm7A [NMR], 1t29A [x-ray, 2.3 Å]			
COMT	271	1-271	100		multiple [96%]	0.337 (-4.68)	0.370 (1.99E-03)
CYP1A2	515	1-515	100		multiple [93%]	0.702 (-0.78)	0.617 (1.39E-04)
CYP2A6	494	1-494	100		multiple [100%]	0.774 (0.07)	0.663 (9.31E-05)
CYP2B6	491	1-491	100		multiple [95%]	0.732 (-0.44)	0.562 (2.30E-04)
CYP2C19	490	1-490	100		multiple [95%]	0.694 (-0.88)	0.621 (1.34E-04)
CYP2C8	490	1-490	100		multiple [100%]	0.765 (-0.03)	0.645 (1.08E-04)
CYP2C9	490	1-490	100		multiple [95%]	0.742 (-0.31)	0.620 (1.35E-04)
CYP2D6	497	1-497	100		multiple [93%]	0.728 (-0.47)	0.615 (1.40E-04)
CYP2E1	493	1-493	100		multiple [94%]	0.736 (-0.38)	0.647 (1.07E-04)
CYP2J2	502	1-502	100		multiple [92%]	0.704 (-0.76)	0.651 (1.03E-04)
CYP3A4	503	1-503	100		multiple [94%]	0.664 (-1.23)	0.594 (1.71E-04)
CYP3A5	502	1-502	100		multiple [93%]	0.690 (-0.93)	0.592 (1.73E-04)
DPYD	1025	1-1025	100		multiple [99%]	0.777 (0.16)	0.223 (2.03E-02)
DRD2	443	30-442	93.2		2rh1A [20.8%]	0.250 (-6.02)	0.359 (2.29E-03)
F5	2224	29-737, 1574-2224	61.2	1y61A [model]			

Supplementary Table B.2 (continued)

Protein	Protein full length	Structure coverage†		Experimental structure	Homology model		
		Residue range	% of sequence coverage	PDB ID [method]	Template PDB [% confident]	QMEAN6 score (Z-score)	ModFOLD4 score (p-value)
G6PD	515	1-515	100		multiple [94%]	0.627 (-1.69)	0.389 (1.55E-03)
GSTP1	210	1-210	100	3dgqA [x-ray, 1.6 Å]			
GSTT1	240	1-240	100		multiple [99%]	0.795 (0.26)	0.637 (1.16E-04)
HMGCR	888	1-870	98.0	1dqaA [x-ray, 2.0 Å]	multiple [64%]	0.424 (-3.74)	0.231 (3.74E-2)
KCNH2	1159	26-135	9.5	4hqaA [x-ray, 1.96 Å]			
KCNH2	1159	667-869	17.5		1q5oA [26.24%]	0.565 (-2.16)	0.491 (4.70E-04)
KCNJ11	390	1-390	100		multiple [84%]	0.430 (-3.95)	0.230 (1.76E-02)
MTHFR	656	57-338	43.0		1zrqC [33.6%]	0.757 (-0.15)	0.370 (1.99E-03)
NQO1	274	1-274	100		multiple [100%]	0.738 (-0.33)	0.569 (2.15E-04)
NR1I2	434	1-434	100		multiple [86%]	0.429 (-4.02)	0.307 (4.88E-03)
P2RY1	373	1-373	100		multiple [96%]	0.352 (-4.78)	0.362 (2.20E-03)
P2RY12	342	1-342	100		multiple [99%]	0.362 (-4.79)	0.459 (6.68E-04)
PTGIS	500	1-500	100		multiple [96%]	0.692 (-0.91)	0.560 (2.34E-04)
PTGS2	604	1-604	100		multiple [91%]	0.742 (-0.24)	0.485 (5.03E-04)
SCN5A	2016	1776-1928	7.6	4dckA [x-ray, 2.2 Å]			
SLC19A1	591	1-591	100		multiple [73%]	0.418 (-3.82)	0.218 (3.10E-02)
SULT1A1	295	1-295	100		multiple [99%]	0.861 (0.95)	0.674 (8.52E-05)
TPMT	245	1-245	100		multiple [93%]	0.718 (-0.55)	0.460 (6.59E-04)
TYMS	313	1-313	100		multiple [97%]	0.735 (-0.45)	0.651 (1.03E-04)
UGT1A1	533	1-533	100		multiple [84%]	0.444 (-3.78)	0.314 (4.36E-03)
VDR	427	1-427	100		multiple [89%]	0.486 (-3.32)	0.320 (4.02E-03)

Supplementary Table B.2 (continued)

†Structure coverage represents the range of protein residues present in the 3D structure and the percentage of structural coverage with respect to the full length of the proteins. Each PDB accession number is indicated by its four letter code followed by the chain ID (in capital letter).

*Details of homology models include the template' PDBs, sequence identity percentage between the target and the template sequences (in case of single-template modeling), or percentage of residues modeled at >90% confident (in case of multi-template modeling). Quality of homology models is evaluated by QMEAN6 and ModFOLD4 scores. QMEAN6 is a reliability score derived from a linear combination of six terms. It estimates model reliability between 0 and 1 (1 represents the best model). QMEAN Z-score for a given model is a comparable quality score to experimental structures of similar size. Models of low quality will have strongly negative QMEAN Z-scores. ModFOLD's global model quality score ranges between 0 and 1. In general, scores greater than 0.4 indicate more complete and confident models, which are highly similar to the native structure. ModFOLD's p-value represents the probability that each model is incorrect.

Table B.3: Complete list of genomic variability data of the 48 VIPs and the statistics of 3D structure maps.

Gene name	All genomic data		Genomic data with available 3D structures		%3D coverage	
	# of functional variants	# of neutral variants	# of functional variants	# of neutral variants	Functional variants	Neutral variants
ABCB1	2	86	0	0	0%	0%
ACE	2	148	0	140	0%	95%
ADH1A	0	17	0	17		100%
ADH1B	1	32	1	32	100%	100%
ADH1C	2	17	2	17	100%	100%
ADRB1	11	20	11	20	100%	100%
ADRB2	4	35	4	35	100%	100%
AHR	2	55	2	18	100%	33%
ALDH1A1	0	31	0	31		100%
ALOX5	16	48	16	48	100%	100%
BRCA1	186	245	58	25	31%	10%
COMT	3	22	3	22	100%	100%
CYP1A2	0	71	0	71		100%
CYP2A6	7	73	7	73	100%	100%
CYP2B6	1	80	1	80	100%	100%
CYP2C19	5	68	5	68	100%	100%
CYP2C8	0	53	0	53		100%
CYP2C9	7	52	7	52	100%	100%
CYP2D6	1	95	1	95	100%	100%
CYP2E1	0	41	0	41		100%
CYP2J2	0	40	0	40		100%
CYP3A4	0	54	0	54		100%
CYP3A5	0	39	0	39		100%
DPYD	4	85	4	85	100%	100%
DRD2	1	28	1	25	100%	89%
F5	8	165	7	91	88%	55%
G6PD	52	38	52	38	100%	100%
GSTP1	2	28	2	28	100%	100%
GSTT1	2	16	2	16	100%	100%
HMGCR	2	25	2	25	100%	100%
KCNH2	117	65	35	9	30%	14%
KCNJ11	51	29	51	29	100%	100%
MTHFR	18	48	8	19	44%	40%
NQO1	1	16	1	16	100%	100%
NR1I2	0	52	0	52		100%

Supplementary Table B.3 (continued)

Gene name	All genomic data		Genomic data with available 3D structures		%3D coverage	
	# of functional variants	# of neutral variants	# of functional variants	# of neutral variants	Functional variants	Neutral variants
P2RY1	0	10	0	10		100%
P2RY12	2	18	2	18	100%	100%
PTGIS	0	61	0	61		100%
PTGS2	2	21	2	21	100%	100%
SCN5A	180	147	18	13	10%	9%
SLC19A1	0	52	0	52		100%
SLCO1B1	2	74	0	0	0%	0%
SULT1A1	1	31	1	31	100%	100%
TPMT	7	16	7	16	100%	100%
TYMS	0	4	0	4		100%
UGT1A1	42	38	42	38	100%	100%
VDR	16	43	16	43	100%	100%
VKORC1	19	5	0	0	0%	0%
Total	779	2537	371	1811	-	-
Average	16	54	8	38	80%	84%
Range	0-186	4-245	0-58	0-140	0-100%	0-100%
# of proteins with variant data	35	48	31	45	-	-
# of proteins whose structure covers > 70% of total variants	-	-	-	-	27	39

Table B.4: Complete list of domain names and relative abundances of functional and neutral variants.

Domain	Domain name	Domain	Domain name
MF_00008	Thymidylate synthase	PR01666	Voltage gated sodium channel, alpha-5 subunit
MF_00812	Thiopurine S-methyltransferase	PR01683	Cytochrome P450, E-class, group I, CYP1
MF_00966	Glucose-6-phosphate dehydrogenase	PR01684	Cytochrome P450, E-class, group I, CYP2A-like
PF00001	G protein-coupled receptor, rhodopsin-like	PR01685	Cytochrome P450, E-class, group I, CYP2B-like
PF00010	Myc-type, basic helix-loop-helix (bHLH) domain	PR01686	Cytochrome P450, E-class, group I, CYP2D-like
PF00027	Cyclic nucleotide-binding domain	PR01687	Cytochrome P450, E-class, group I, CYP2E-like
PF00043	Glutathione S-transferase, C-terminal	PR01688	Cytochrome P450, E-class, group I, CYP2J-like
PF00067	Cytochrome P450	PR01689	Cytochrome P450, E-class, CYP3A
PF00097	Zinc finger, C3HC4 RING-type	PS00059	Alcohol dehydrogenase, zinc-type, conserved site
PF00104	Nuclear hormone receptor, ligand-binding, core	PS00079	Multicopper oxidase, copper-binding site
PF00105	Zinc finger, nuclear hormone receptor-type	PS00081	Lipoxygenase, conserved site
PF00107	Alcohol dehydrogenase, C-terminal	PS00086	Cytochrome P450, conserved site
PF00171	Aldehyde dehydrogenase domain	PS00198	4Fe-4S ferredoxin, iron-sulphur binding, conserved site
PF00201	UDP-glucuronosyl/UDP-glucosyltransferase	PS00518	Zinc finger, RING-type, conserved site
PF00479	Glucose-6-phosphate dehydrogenase, NAD-binding	PS00711	Lipoxygenase, iron binding site
PF00520	Ion transport domain	PS50089	Zinc finger, RING-type
PF00685	Sulfotransferase domain	PS50112	PAS domain
PF00754	Coagulation factor 5/8 C-terminal type domain	PS50113	PAS-associated, C-terminal
PF00989	PAS fold	PS50156	Sterol-sensing domain
PF01007	Potassium channel, inwardly rectifying, Kir	PS50262	GPCR, rhodopsin-like, 7TM
PF01180	Dihydroorotate dehydrogenase, class 1/ 2	PS50405	Glutathione S-transferase, C-terminal-like
PF01401	Peptidase M2, peptidyl-dipeptidase A	PS51379	4Fe-4S ferredoxin-type, iron-sulphur binding domain
PF01477	PLAT/LH2 domain	PS51393	Lipoxygenase, C-terminal
PF01596	O-methyltransferase, family 3	PS51557	Catechol O-methyltransferase
PF01770	Reduced folate carrier	PTHR10572	Hydroxymethylglutaryl-CoA reductase, class I/II
PF02219	Methylenetetrahydrofolate reductase	PTHR11695	Alcohol dehydrogenase superfamily, zinc-type
PF02525	Flavodoxin-like fold	PTHR11771	Lipoxygenase
PF02781	Glucose-6-phosphate dehydrogenase, C-terminal	PTHR24231:SF2	P2Y1 purinoceptor
PF02798	Glutathione S-transferase, N-terminal	PTHR24233:SF0	P2Y12 purinoceptor

Supplementary Table B.4 (continued)

Domain	Domain name	Domain	Domain name
PF03098	Haem peroxidase, animal	PTHR24248	Adrenoceptor family
PF05724	TPMT family	PTHR24248:SF13	Dopamine D2 receptor
PF06512	Sodium ion transport-associated	PTHR24248:SF21	Beta 2 adrenoceptor
PF07732	Multicopper oxidase, type 3	PTHR24248:SF3	Beta 1 adrenoceptor
PF08240	Alcohol dehydrogenase GroES-like	SM00015	IQ motif, EF-hand binding site
PF08447	PAS fold-3	SM00086	PAC motif
PF11933	Domain of unknown function DUF3451	SSF103473	Major facilitator superfamily domain, general substrate transporter
PF12820	BRCA1, serine-rich domain	SSF46548	Alpha-helical ferredoxin
PF14691	Dihydroprymidine dehydrogenase domain II	SSF48113	Haem peroxidase
PIRSF000047	Cytochrome P450, cholesterol 7-alpha-monooxygenase-type	SSF48508	Nuclear hormone receptor, ligand-binding
PIRSF000354	Coagulation factor 5/8	SSF49503	Cupredoxin
PIRSF001734	Breast cancer type 1 susceptibility protein (BRCA1)	SSF49723	Lipase/lipooxygenase, PLAT/LH2
PIRSF037177	Catechol O-methyltransferase, eukaryotic	SSF49785	Galactose-binding domain-like
PIRSF500628	Prostacyclin synthase	SSF50129	GroES (chaperonin 10)-like
PIRSF500793	Folate transporter 1	SSF51206	Cyclic nucleotide-binding-like
PR00170	Voltage gated sodium channel, alpha subunit	SSF52113	BRCT domain
PR00242	Dopamine receptor family	SSF52540	P-loop containing nucleoside triphosphate hydrolase
PR00350	Vitamin D receptor	SSF52833	Thioredoxin-like fold
PR00398	Steroid hormone receptor	SSF53720	Aldehyde/histidinol dehydrogenase
PR00457	Haem peroxidase, animal, subgroup	SSF55035	Hydroxymethylglutaryl-CoA reductase, class I/II, NAD/NADP-binding
PR00463	Cytochrome P450, E-class, group I	SSF55831	Thymidylate synthase/dCMP hydroxymethylase domain
PR00464	Cytochrome P450, E-class, group II	SSF56542	Hydroxymethylglutaryl-CoA reductase, class I/II, substrate-binding
PR00465	Cytochrome P450, E-class, group IV	SSF81296	Immunoglobulin E-set
PR00467	Lipoxygenase, mammalian	TIGR00533	Hydroxymethylglutaryl-CoA reductase, eukaryotic/arcaheal type
PR01268	Glutathione S-transferase, Pi class	TIGR00677	Eukaryotic-type methylenetetrahydrofolate reductase
PR01332	Potassium channel, inwardly rectifying, Kir6.2	TIGR00920	Hydroxymethylglutaryl-CoA reductase, metazoan
PR01463	Potassium channel, voltage-dependent, EAG/ELK/ERG	TIGR01037	Dihydroorotate dehydrogenase domain
PR01470	Potassium channel, voltage-dependent, ERG		

Supplementary Table B.4 (continued)

* The prefix of Domain ID indicates the data source. Abbreviations are MF: HAMAP; PF: Pfam; PIRSF: PIRSF; PR: PRINTS, PS: PROSITE; PTHR: PANTHER; SM: SMART; SSF: SUPERFAMILY; TIGR: TIGRFAMS.

Table B.5: Complete statistics of Fisher's exact test for enriched structural features present in functional and neutral mutations.

Structural features		Indicators	Descriptions	Fisher's exact test (one-tailed)
Inter-residue bonding	Disulfide bond	At Cys	Wild type residue = Cys	0.0003*
		Loss of disulfide bond	Wild type residue forms disulfide bond (when wile type AA = Cys) and mutant residue eliminates the bond	0.4039
	Salt bridge	At charged AA	Wild type residue = charged AA	0.6537
		At salt bridge	Wild type residue forms salt bridges (when wild type AA = charged residues)	0.9081
		Loss of salt bridge	Mutant residues eliminate salt bridges (changes to opposite charged/neutral AA)	0.5021
Protein stability	Key stabilizing residues	At SC	Wild type residue = a stabilization center (SC)	0.0848
		At SR	Wild type residue = a stabilizing residue (SR)	0.0037*
	Stability change	Any change of stability	Mutant residue reduces or increases stability ($\Delta\Delta G \geq 0.5$ or ≤ -0.5 kcal/mol)	0.0311*
		Destabilizing	Mutant residue reduces stability ($\Delta\Delta G \geq 0.5$ kcal/mol)	0.0588
		Highly destabilizing	Mutant residue highly reduces stability ($\Delta\Delta G \geq 4$ kcal/mol)	0.4059
		Stabilizing	Mutant residue increases stability ($\Delta\Delta G \leq -0.5$ kcal/mol)	0.1317
		Highly stabilizing	Mutant residue highly increases stability ($\Delta\Delta G \leq -2$ kcal/mol)	none
Protein flexibility	Interdomain	At hinge	Located at hinge site	0.6563
	From crystallography	Highly rigid	Is a highly rigid residue (B-factorNorm ≤ -0.523)	0.6490
		Highly dynamic	Is a highly dynamic residue (B-factorNorm ≥ 1.0909)	0.7997
		Highly rigid or highly dynamic	Highly rigid or highly dynamic residue (B-factorNorm ≤ -0.523 or ≥ 1.0909)	0.7776
	From MD simulations	Conformationally rigid	Located at conformationally rigid site (FlexPred label = rigid)	0.0218*
		Highly rigid	Is a highly rigid residue (RMSFNorm ≤ -0.4609)	none
		Highly dynamic	Is a highly dynamic residue (RMSFNorm ≥ 0.964)	0.1088
		Highly rigid or highly dynamic	Highly rigid or highly dynamic residue (RMSFNorm ≤ -0.4609 or ≥ 0.964)	0.1088

Supplementary Table B.5 (continued)

Structural features		Indicators	Descriptions	Fisher's exact test (one-tailed)
Drug binding capability	Binding site predictions	At binding site	Located at binding site	0.0003*
		At 5 Å of binding site	Located within 5 Å of binding site (Cα-Cα distance)	0.0123*
		At 10 Å of binding site	Located within 10 Å of binding site (Cα-Cα distance)	0.0005*
	Catalytic site predictions	Non-optimal residue	Is a nonoptimal residue ($\Gamma < 0$ kcal/mol)	0.2014
		Highly-non optimal residue	Is a highly nonoptimal residue ($\Gamma < -3.47$ kcal/mol), most likely catalytic sites	0.1738
	Amino acid change at binding site	Change of hydropathy, at binding site	Induced hydropathy change (hydrophobic ←→ hydrophilic) when located at binding sites	0.2854
		Change of volume, at binding site	Induced volume change (very large to small/very small, very small to large/very large) when located at binding sites	0.2241
		Change of charge, at binding site	Induced charge change (positive ←→ negative) when located at binding sites	0.4870
		Change of H-bond, at binding site	Induced H-bond change (H-bond donor ←→ acceptor) when located at binding sites	0.4870
		Change of any kind, at binding site	Induced any of the four changes when located at binding sites	0.0954
	Amino acid change around 5 Å of binding sites	Change of hydropathy, around 5 Å of binding sites	Induced hydropathy change when located around 5 Å of binding sites	0.4943
		Change of volume, around 5 Å of binding sites	Induced volume change when located around 5 Å of binding sites	0.0385*
		Change of charge, around 5 Å of binding sites	Induced charge change when located around 5 Å of binding sites	0.6429
		Change of H-bond, around 5 Å of binding sites	Induced H-bond change when located around 5 Å of binding sites	0.6429
		Change of any kind, around 5 Å of binding sites	Induced any of the four changes when located around 5 Å of binding sites	0.0803
	Amino acid change around 10 Å of binding sites	Change of hydropathy, around 10 Å of binding sites	Induced hydropathy change when located around 10 Å of binding sites	0.1117
		Change of volume, around 10 Å of binding sites	Induced volume change when located around 10 Å of binding sites	0.0046*
		Change of charge, around 10 Å of binding sites	Induced charge change when located around 10 Å of binding sites	0.9107
		Change of H-bond, around 10 Å of binding sites	Induced H-bond change when located around 10 Å of binding sites	0.9585
		Any changes around 10 Å of binding sites	Induced any of the four changes when located around 10 Å of binding sites	0.0105*

Supplementary Table B.5 (continued)

Structural features		Indicators	Descriptions	Fisher's exact test (one-tailed)
Protein-protein interaction	Interaction site	At patch	Located on protein patch	< 0.0001*
Residue localization	Unusual AA	At Gly/Pro	Induced Gly/Pro change	0.0015*
	Residue localization	At domain	Located in protein domains	0.2355
	Residue localization	At structural site	Located at structural site	0.9978
	Residue localization	At buried site	Locate in buried site (RSA ≤ 20%)	0.0533
	Residue localization	At core	Located at the core region (RSA ≤ 5%)	0.0021*
Amino acid dissimilarity	Grantham score, by location	Large AA dissimilarity at any site	Induced large amino acid change (Grantham score ≥ 100) when located at any sites	< 0.0001*
		Large AA dissimilarity at binding site	Induced large amino acid change (Grantham score ≥ 100) when located at binding sites	0.0259*
		Large AA dissimilarity at within 5 Å of binding sites	Induced large amino acid change (Grantham score ≥ 100) when located within 5 Å of binding sites	0.0079*
		Large AA dissimilarity at within 10 Å of binding sites	Induced large amino acid change (Grantham score ≥ 100) when located within 10 Å of binding sites	< 0.0001*
		Large AA dissimilarity at domain	Induced large amino acid change (Grantham score ≥ 100) when located in protein domains	< 0.0001*
		Large AA dissimilarity at any domain boundary	Induced large amino acid change (Grantham score ≥ 100) when located in domain boundaries	0.0045*
		Large AA dissimilarity at structural site	Induced large amino acid change (Grantham score ≥ 100) when located at structural sites	< 0.0001*
		Large AA dissimilarity at non-structural sites	Induced large amino acid change (Grantham score ≥ 100) when located at non-structural sites	0.6145

Supplementary Table B.5 (continued)

Structural features		Indicators	Descriptions	Fisher's exact test (one-tailed)
Amino acid secondary structure preference	Secondary structure break	Any type	Induced secondary structure break (preferred to least preferred amino acid propensity) at any secondary structure types	0.1674
		Coil break	Induced secondary structure break (preferred to least preferred amino acid propensity) at coil	0.6199
		Strand break	Induced secondary structure break (preferred to least preferred amino acid propensity) at strand	0.2454
		3-turn helix break	Induced secondary structure break (preferred to least preferred amino acid propensity) at 3-turn helix	0.4586
		α-helix break	Induced secondary structure break (preferred to least preferred amino acid propensity) at α-helix	0.1256
		bend break	Induced secondary structure break (preferred to least preferred amino acid propensity) at bend	0.5348
		turn break	Induced secondary structure break (preferred to least preferred amino acid propensity) at turn	0.4898

Significant p-values (α = .05) are designated with '*'. Non-significant p-values indicate Prob(testing features) is greater for functional variants.

REFERENCES

1. Kumar, S., et al., *Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations.* Trends Genet, 2011. 27(9): p. 377-86.

2. MacArthur, D.G., et al., *A systematic survey of loss-of-function variants in human protein-coding genes.* Science, 2012. 335(6070): p. 823-8.

3. Pelak, K., et al., *The characterization of twenty sequenced human genomes.* PLoS Genet, 2010. 6(9).

4. Teo, S.M., et al., *A population-based study of copy number variants and regions of homozygosity in healthy Swedish individuals.* J Hum Genet, 2011. 56(7): p. 524-33.

5. Cooper, D.N., et al., *Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics.* Hum Mutat, 2010. 31(6): p. 631-55.

6. Lyon, G.J. and K. Wang, *Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress.* Genome Med, 2012. 4(7): p. 58.

7. Green, R.C., H.L. Rehm, and I.S. Kohane, *Challenges in the Clinical Use of Genome Sequencing*, in *Genomic and personalized medicine*, G.S. Ginsburg and H.F. Willard, Editors. 2012, Elsevier/Academic Press,: London. p. 1 online resource (2 v. (xxxv, 1305 p.)).

8. Cargill, M., et al., *Characterization of single-nucleotide polymorphisms in coding regions of human genes.* Nat Genet, 1999. 22(3): p. 231-8.

9. Spencer, D.H., K.L. Bubb, and M.V. Olson, *Detecting disease-causing mutations in the human genome by haplotype matching.* Am J Hum Genet, 2006. 79(5): p. 958-64.

10. Storz, J.F. and A.J. Zera, *Experimental approaches to evaluate the contributions of candidate protein-coding mutations to phenotypic evolution.* Methods Mol Biol, 2011. 772: p. 377-96.

11. Capriotti, E., et al., *Bioinformatics for personal genome interpretation.* Brief Bioinform, 2012. 13(4): p. 495-512.

12. Chen, J.J. and B.R. Shen, *Computational Analysis of Amino Acid Mutation: A Proteome Wide Perspective.* Current Proteomics, 2009. 6(4): p. 228-234.

13. Coassin, S., A. Brandstatter, and F. Kronenberg, *Lost in the space of bioinformatic tools: a constantly updated survival guide for genetic epidemiology. The GenEpi Toolbox.* Atherosclerosis, 2010. 209(2): p. 321-35.

14. Cooper, G.M. and J. Shendure, *Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data.* Nat Rev Genet, 2011. 12(9): p. 628-40.

15. Liu, X., X. Jian, and E. Boerwinkle, *dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions.* Hum Mutat, 2011. 32(8): p. 894-9.

16. Kumar, P., S. Henikoff, and P.C. Ng, *Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.* Nat Protoc, 2009. 4(7): p. 1073-81.

17. Chun, S. and J.C. Fay, *Identification of deleterious mutations within three human genomes.* Genome Res, 2009. 19(9): p. 1553-61.

18. Reva, B., Y. Antipin, and C. Sander, *Predicting the functional impact of protein mutations: application to cancer genomics.* Nucleic Acids Res, 2011. 39(17): p. e118.

19. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. 7(4): p. 248-9.

20. Schwarz, J.M., et al., *MutationTaster evaluates disease-causing potential of sequence alterations.* Nat Methods, 2010. 7(8): p. 575-6.

21. Jordan, D.M., V.E. Ramensky, and S.R. Sunyaev, *Human allelic variation: perspective from protein function, structure, and evolution.* Curr Opin Struct Biol, 2010. 20(3): p. 342-50.

22. Gonzalez-Perez, A. and N. Lopez-Bigas, *Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel.* Am J Hum Genet, 2011. 88(4): p. 440-9.

23. Gamazon, E.R., A.D. Skol, and M.A. Perera, *The limits of genome-wide methods for pharmacogenomic testing.* Pharmacogenet Genomics, 2012. 22(4): p. 261-72.

24. Dorfman, R., et al., *Do common in silico tools predict the clinical consequences of amino-acid substitutions in the CFTR gene?* Clin Genet, 2010. 77(5): p. 464-73.

25. Giudicessi, J.R., et al., *Phylogenetic and physicochemical analyses enhance the classification of rare nonsynonymous single nucleotide variants in type 1 and 2 long-QT syndrome.* Circ Cardiovasc Genet, 2012. 5(5): p. 519-28.

26. Dewey, F.E., et al., *DNA sequencing: clinical applications of new DNA sequencing technologies.* Circulation, 2012. 125(7): p. 931-44.

27. Hicks, S., et al., *Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed.* Hum Mutat, 2011. 32(6): p. 661-8.

28. Ng, P.C. and S. Henikoff, *Predicting the effects of amino acid substitutions on protein function.* Annu Rev Genomics Hum Genet, 2006. 7: p. 61-80.

29. Li, M.X., et al., *A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases.* Nucleic Acids Res, 2012. 40(7): p. e53.

30. Wu, J. and R. Jiang, *Prediction of deleterious nonsynonymous single-nucleotide polymorphism for human diseases.* ScientificWorldJournal, 2013. 2013: p. 675851.

31. Kanchi, K.L., et al., *Integrated analysis of germline and somatic variants in ovarian cancer.* Nat Commun, 2014. 5: p. 3156.

32. Ohanian, M., R. Otway, and D. Fatkin, *Heuristic methods for finding pathogenic variants in gene coding sequences.* J Am Heart Assoc, 2012. 1(5): p. e002642.

33. Teng, S., E. Michonova-Alexova, and E. Alexov, *Approaches and resources for prediction of the effects of non-synonymous single nucleotide polymorphism on protein function and interactions.* Curr Pharm Biotechnol, 2008. 9(2): p. 123-33.

34. Verma, R., U. Schwaneberg, and D. Roccatano, *Computer-aided protein directed evolution: a review of web servers, databases and other computational tools for protein engineering.* Computational and Structural Biotechnology Journal, 2012. 2(3): p. e201209008.

35. Capriotti, E. and R.B. Altman, *Improving the prediction of disease-related variants using protein three-dimensional structure.* BMC Bioinformatics, 2011. 12 Suppl 4: p. S3.

36. David, A., et al., *Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs.* Hum Mutat, 2012. 33(2): p. 359-63.

37. Yue, W.W., D.S. Froese, and P.E. Brennan, *The role of protein structural analysis in the next generation sequencing era.* Top Curr Chem, 2014. 336: p. 67-98.

38. Izarzugaza, J.M.G., et al., *wKinMut: An integrated tool for the analysis and interpretation of mutations in human protein kinases.* Bmc Bioinformatics, 2013. 14.

39. Jordan, D.M., et al., *Development and validation of a computational method for assessment of missense variants in hypertrophic cardiomyopathy.* Am J Hum Genet, 2011. 88(2): p. 183-92.

40. Preeprem, T. and G. Gibson, *An association-adjusted consensus deleterious scheme to classify homozygous Mis-sense mutations for personal genome interpretation.* BioData Min, 2013. 6(1): p. 24.

41. Ng, P.C., et al., *Genetic variation in an individual human exome.* PLoS Genet, 2008. 4(8): p. e1000160.

42. Xue, Y., et al., *Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing.* Am J Hum Genet, 2012. 91(6): p. 1022-32.

43. Luu, T.D., et al., *MSV3d: database of human MisSense Variants mapped to 3D protein structure.* Database (Oxford), 2012. 2012: p. bas018.

44. Mottaz, A., et al., *Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar.* Bioinformatics, 2010. 26(6): p. 851-2.

45. Hindorff L.A., et al. *A Catalog of Published Genome-Wide Association Studies*. March 28, 2013; Available from: www.genome.gov/gwastudies.

46. Patel, C.J., et al., *Whole genome sequencing in support of wellness and health maintenance.* Genome Med, 2013. 5(6): p. 58.

47. Habegger, L., et al., *VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment.* Bioinformatics, 2012. 28(17): p. 2267-9.

48. Derrien, T., et al., *The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression.* Genome Res, 2012. 22(9): p. 1775-89.

49. National Center for Biotechnology Information. *Database of Single Nucleotide Polymorphisms (dbSNP)*. 2013; Available from: http://www.ncbi.nlm.nih.gov/SNP/.

50. NHLBI Exome Sequencing Project. *Exome Variant Server*. 2012 August 30; Available from: http://evs.gs.washington.edu/EVS/.

51. UniProt Consortium, *Reorganizing the protein space at the Universal Protein Resource (UniProt).* Nucleic Acids Res, 2012. 40(Database issue): p. D71-5.

52. McKusick-Nathans Institute of Genetic Medicine and Johns Hopkins University. *Online Mendelian Inheritance in Man, OMIM®*. April 1, 2013; Available from: http://omim.org/.

53. Davydov, E.V., et al., *Identifying a high fraction of the human genome to be under selective constraint using GERP++*. PLoS Comput Biol, 2010. 6(12): p. e1001025.

54. Pollard, K.S., et al., *Detection of nonneutral substitution rates on mammalian phylogenies.* Genome Res, 2010. 20(1): p. 110-21.

55. Lindblad-Toh, K., et al., *A high-resolution map of human evolutionary constraint using 29 mammals.* Nature, 2011. 478(7370): p. 476-82.

56. Grantham, R., *Amino acid difference formula to help explain protein evolution.* Science, 1974. 185(4154): p. 862-4.

57. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Res, 2010. 38(16): p. e164.

58. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* Fly (Austin), 2012. 6(2): p. 80-92.

59. Xie, L. and P.E. Bourne, *Functional coverage of the human genome by existing structures, structural genomics targets, and homology models.* PLoS Comput Biol, 2005. 1(3): p. e31.

60. Bernstein, F.C., et al., *The Protein Data Bank: a computer-based archival file for macromolecular structures.* J Mol Biol, 1977. 112(3): p. 535-42.

61. Haas, J., et al., *The Protein Model Portal--a comprehensive resource for protein structure and model information.* Database (Oxford), 2013. 2013: p. bat031.

62. Krieger, E., et al., *Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8.* Proteins, 2009. 77 Suppl 9: p. 114-22.

63. Benkert, P., T. Schwede, and S.C. Tosatto, *QMEANclust: estimation of protein model quality by combining a composite scoring function with structural density information.* BMC Struct Biol, 2009. 9: p. 35.

64. McGuffin, L.J., M.T. Buenavista, and D.B. Roche, *The ModFOLD4 server for the quality assessment of 3D protein models.* Nucleic Acids Res, 2013.

65. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem, 2004. 25(13): p. 1605-12.

66. Worth, C.L., R. Preissner, and T.L. Blundell, *SDM--a server for predicting effects of mutations on protein stability and malfunction.* Nucleic Acids Res, 2011. 39(Web Server issue): p. W215-22.

67. Smith, R.E., et al., *Andante: reducing side-chain rotamer search space during comparative modeling using environment-specific substitution probabilities.* Bioinformatics, 2007. 23(9): p. 1099-105.

68. Capriotti, E., P. Fariselli, and R. Casadio, *I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure.* Nucleic Acids Res, 2005. 33(Web Server issue): p. W306-10.

69. Dosztanyi, Z., et al., *SCide: identification of stabilization centers in proteins.* Bioinformatics, 2003. 19(7): p. 899-900.

70. Dosztanyi, Z., A. Fiser, and I. Simon, *Stabilization centers in proteins: identification, characterization and predictions.* J Mol Biol, 1997. 272(4): p. 597-612.

71. Magyar, C., et al., *SRide: a server for identifying stabilizing residues in proteins.* Nucleic Acids Res, 2005. 33(Web Server issue): p. W303-5.

72. Wass, M.N., L.A. Kelley, and M.J. Sternberg, *3DLigandSite: predicting ligand-binding sites using similar structures.* Nucleic Acids Res, 2010. 38(Web Server issue): p. W469-73.

73. Nimrod, G., et al., *Detection of functionally important regions in "hypothetical proteins" of known structure.* Structure, 2008. 16(12): p. 1755-63.

74. Nimrod, G., et al., *In silico identification of functional regions in proteins.* Bioinformatics, 2005. 21 Suppl 1: p. i328-37.

75. de Brevern, A.G., et al., *PredyFlexy: flexibility and local structure prediction from sequence.* Nucleic Acids Res, 2012. 40(Web Server issue): p. W317-22.

76. Kuznetsov, I.B. and M. McDuffie, *FlexPred: a web-server for predicting residue positions involved in conformational switches in proteins.* Bioinformation, 2008. 3(3): p. 134-6.

77. Kuznetsov, I.B., *Ordered conformational change in the protein backbone: prediction of conformationally variable positions from sequence and low-resolution structural data.* Proteins, 2008. 72(1): p. 74-87.

78. Reimand, J., T. Arak, and J. Vilo, *g:Profiler--a web server for functional interpretation of gene lists (2011 update).* Nucleic Acids Res, 2011. 39(Web Server issue): p. W307-15.

79. Stark, C., et al., *The BioGRID Interaction Database: 2011 update.* Nucleic Acids Res, 2011. 39(Database issue): p. D698-704.

80. Hakenberg, J., et al., *A SNPshot of PubMed to associate genetic variants with drugs, diseases, and adverse reactions.* J Biomed Inform, 2012. 45(5): p. 842-50.

81. Ireland, J., et al., *Large-scale characterization of public database SNPs causing non-synonymous changes in three ethnic groups.* Hum Genet, 2006. 119(1-2): p. 75-83.

82. Cmarik, J.L., *From bioinformatics to bioassays: gleaning insights into protein structure-function from disease-associated nsSNPs.* Mol Interv, 2008. 8(4): p. 162-4.

83. Chasman, D.I., *Functional assessment of amino acid variation caused by single nucleotide polymorphisms: a structural view*, in *Protein structure : determination, analysis, and applications for drug discovery*, D.I. Chasman, Editor. 2003, Marcel Dekker: New York. p. xiv, 606 p.

84. Federoff, H.J., *Alzheimer's disease: reducing the burden with ApoE2.* Gene Ther, 2005. 12(13): p. 1019-29.

85. Corder, E.H., et al., *Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families.* Science, 1993. 261(5123): p. 921-3.

86. Reiman, E.M., et al., *Preclinical evidence of Alzheimer's disease in persons homozygous for the epsilon 4 allele for apolipoprotein E.* N Engl J Med, 1996. 334(12): p. 752-8.

87. Breslow, J.L., et al., *Studies of familial type III hyperlipoproteinemia using as a genetic marker the apoE phenotype E2/2.* J Lipid Res, 1982. 23(8): p. 1224-35.

88. Weisgraber, K.H., *Apolipoprotein E: structure-function relationships.* Adv Protein Chem, 1994. 45: p. 249-302.

89. Bolino, A., et al., *Charcot-Marie-Tooth type 4B is caused by mutations in the gene encoding myotubularin-related protein-2.* Nat Genet, 2000. 25(1): p. 17-9.

90. Bolino, A., et al., *Denaturing high-performance liquid chromatography of the myotubularin-related 2 gene (MTMR2) in unrelated patients with Charcot-Marie-Tooth disease suggests a low frequency of mutation in inherited neuropathy.* Neurogenetics, 2001. 3(2): p. 107-9.

91. Office of Communications and Public Liaison, National Institute of Neurological Disorders and Stroke, and National Institutes of Health. *Charcot-Marie-Tooth Disease Fact Sheet*. February 15, 2011 [cited 2013 May 19]; Available from: http://www.ninds.nih.gov/disorders/charcot_marie_tooth/detail_charcot_marie_tooth.htm.

92. Zwicker, J.I., et al., *The thrombospondin-1 N700S polymorphism is associated with early myocardial infarction without altering von Willebrand factor multimer size.* Blood, 2006. 108(4): p. 1280-3.

93. Carlson, C.B., et al., *Influences of the N700S thrombospondin-1 polymorphism on protein structure and stability.* J Biol Chem, 2008. 283(29): p. 20069-76.

94. Granados-Riveron, J.T., et al., *Alpha-cardiac myosin heavy chain (MYH6) mutations affecting myofibril formation are associated with congenital heart defects.* Hum Mol Genet, 2010. 19(20): p. 4007-16.

95. Carniel, E., et al., *Alpha-myosin heavy chain: a sarcomeric gene associated with dilated and hypertrophic phenotypes of cardiomyopathy.* Circulation, 2005. 112(1): p. 54-9.

96. Rydberg, E.H., et al., *Mechanistic analyses of catalysis in human pancreatic alpha-amylase: detailed kinetic and structural studies of mutants of three conserved carboxylic acids.* Biochemistry, 2002. 41(13): p. 4492-502.

97. Numao, S., et al., *Probing the role of the chloride ion in the mechanism of human pancreatic alpha-amylase.* Biochemistry, 2002. 41(1): p. 215-25.

98. Petsko, G.A. and D. Ringe, *From sequence to structure*, in *Protein Structure and Function*, E. Lawrence and M. Robertson, Editors. 2004, New Science Press: London. p. 1-49.

99. Betts, M.J. and R.B. Russell, *Amino Acid Properties and Consequences of Substitutions*, in *Bioinformatics for geneticists*, M.R. Barnes and I.C. Gray, Editors. 2003, Wiley: New Jersey. p. 289-316.

100. Brayer, G.D., Y. Luo, and S.G. Withers, *The structure of human pancreatic alpha-amylase at 1.8 A resolution and comparisons with related enzymes.* Protein Sci, 1995. 4(9): p. 1730-42.

101. Kelley, L.A. and M.J. Sternberg, *Protein structure prediction on the Web: a case study using the Phyre server.* Nat Protoc, 2009. 4(3): p. 363-71.

102. Preeprem, T. and G. Gibson, *AACDS—a database for personal genome interpretation.* BMC Genomics, 2014 (in preparation).

103. Kim, J.I., et al., *A highly annotated whole-genome sequence of a Korean individual.* Nature, 2009. 460(7258): p. 1011-5.

104. Hu, H., et al., *VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix.* Genet Epidemiol, 2013. 37(6): p. 622-34.

105. Sifrim, A., et al., *eXtasy: variant prioritization by genomic data fusion.* Nat Methods, 2013. 10(11): p. 1083-4.

106. A. Javed, S. Agrawal, and P.C. Ng., *Phen-Gen: Combining Phenotype and Genotype to Predict Causal Variants in Rare Disorders.* (under reviewed), 2014.

107. Hindorff, L.A., et al. *A Catalog of Published Genome-Wide Association Studies*. March 28, 2013 [cited 2013 October 10]; Available from: www.genome.gov/gwastudies.

108. Tokuriki, N. and D.S. Tawfik, *Stability effects of mutations and protein evolvability.* Curr Opin Struct Biol, 2009. 19(5): p. 596-604.

109. National Center for Biotechnology Information. *Database of Single Nucleotide Polymorphisms (dbSNP).* 2013 [cited 2014 January 16]; Available from: http://www.ncbi.nlm.nih.gov/SNP/.

110. Preeprem, T. and G. Gibson, *SDS, a structural disruption score for assessment of missense variant deleteriousness.* Frontiers in Statistical Genetics and Methodology, 2014. 5(82).

111. Heinzen, E.L., et al., *Exome sequencing followed by large-scale genotyping fails to identify single rare variants of large effect in idiopathic generalized epilepsy.* Am J Hum Genet, 2012. 91(2): p. 293-302.

112. Noebels, J.L., *The biology of epilepsy genes.* Annu Rev Neurosci, 2003. 26: p. 599-625.

113. Motono, C., et al., *SAHG, a comprehensive database of predicted structures of all human proteins.* Nucleic Acids Res, 2011. 39(Database issue): p. D487-93.

114. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* J Mol Biol, 1993. 234(3): p. 779-815.

115. Quevillon, E., et al., *InterProScan: protein domains identifier.* Nucleic Acids Res, 2005. 33(Web Server issue): p. W116-20.

116. Kryshtafovych, A., et al., *Assessment of the assessment: Evaluation of the model quality estimates in CASP10.* Proteins, 2013.

117. Gunasekaran, K., et al., *Stereochemical punctuation marks in protein structures: glycine and proline containing helix stop signals.* J Mol Biol, 1998. 275(5): p. 917-32.

118. Darby, N. and T.E. Creighton, *Disulfide bonds in protein folding and stability.* Methods Mol Biol, 1995. 40: p. 219-52.

119. Ferre, F. and P. Clote, *DiANNA 1.1: an extension of the DiANNA web server for ternary cysteine classification.* Nucleic Acids Res, 2006. 34(Web Server issue): p. W182-5.

120. Dehouck, Y., et al., *PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality.* BMC Bioinformatics, 2011. 12: p. 151.

121. Khan, S. and M. Vihinen, *Performance of protein stability predictors.* Hum Mutat, 2010. 31(6): p. 675-84.

122. Teilum, K., J.G. Olsen, and B.B. Kragelund, *Functional aspects of protein flexibility.* Cell Mol Life Sci, 2009. 66(14): p. 2231-47.

123. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. 22(12): p. 2577-637.

124. Petrovski, S., et al., *Genic intolerance to functional variation and the interpretation of personal genomes.* PLoS Genet, 2013. 9(8): p. e1003709.

125. Kitamura, K., et al., *Three human ARX mutations cause the lissencephaly-like and mental retardation with epilepsy-like pleiotropic phenotypes in mice.* Hum Mol Genet, 2009. 18(19): p. 3708-24.

126. Olsen, R.W., *Combining Ubiquitin Deficiency and GABA-Mediated Inhibition Equals Seizures?* Epilepsy Curr, 2011. 11(3): p. 96-8.

127. Cole-Edwards, K.K. and N.G. Bazan, *Lipid signaling in experimental epilepsy.* Neurochem Res, 2005. 30(6-7): p. 847-53.

128. Kuebler, D., et al., *Genetic suppression of seizure susceptibility in Drosophila.* J Neurophysiol, 2001. 86(3): p. 1211-25.

129. Xu, H.S., et al., *Aligning protein sequence and analysing substitution pattern using a class-specific matrix.* J Biosci, 2010. 35(2): p. 295-314.

130. Guler, G.D., et al., *Human DNA helicase B (HDHB) binds to replication protein A and facilitates cellular recovery from replication stress.* J Biol Chem, 2012. 287(9): p. 6469-81.

131. Lonnqvist, T., et al., *Recessive twinkle mutations cause severe epileptic encephalopathy.* Brain, 2009. 132(Pt 6): p. 1553-62.

132. Voudris, K.A., et al., *Serum amylase, pancreatic amylase and lipase concentrations in epileptic children treated with carbamazepine monotherapy.* Clin Chim Acta, 2004. 350(1-2): p. 175-80.

133. Voudris, K., et al., *Serum total amylase, pancreatic amylase and lipase activities in epileptic children treated with sodium valproate monotherapy.* Brain Dev, 2006. 28(9): p. 572-5.

134. Guerrero, R., et al., *A PTG variant contributes to a milder phenotype in Lafora disease.* PLoS One, 2011. 6(6): p. e21294.

135. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. 28(1): p. 27-30.

136. Lukasiuk, K. and A. Pitkanen, *Molecular basis of acquired epileptogenesis.* Handb Clin Neurol, 2012. 107: p. 3-12.

137. Garofalo, S., M. Cornacchione, and A. Di Costanzo, *From genetics to genomics of epilepsy.* Neurol Res Int, 2012. 2012: p. 876234.

138. Cavalleri, G.L. and N. Delanty, *Opportunities and Challenges for Genome Sequencing in the Clinic.* Challenges and Opportunities of Next-Generation Sequencing for Biomedical Research, 2012. 89: p. 65-83.

139. Ferraro, T.N., et al., *Strategies for Studying the Epilepsy Genome*, in *Jasper's Basic Mechanisms of the Epilepsies*, J.L. Noebels, et al., Editors. 2012: Bethesda (MD).

140. Yandell, M., et al., *A probabilistic disease-gene finder for personal genomes.* Genome Res, 2011. 21(9): p. 1529-42.

141. Robinson, P.N., et al., *Improved exome prioritization of disease genes through cross-species phenotype comparison.* Genome Research, 2014. 24(2): p. 340-348.

142. Jia, P., J.M. Ewers, and Z. Zhao, *Prioritization of epilepsy associated candidate genes by convergent analysis.* PLoS One, 2011. 6(2): p. e17162.

143. Epi4K Consortium and Epilepsy PhenomeGenome Project, *De novo mutations in epileptic encephalopathies.* Nature, 2013. 501(7466): p. 217-21.

144. Dybowski, J.N., et al., *Improved Bevirimat resistance prediction by combination of structural and sequence-based classifiers.* BioData Min, 2011. 4: p. 26.

145. Izarzugaza, J.M., et al., *wKinMut: an integrated tool for the analysis and interpretation of mutations in human protein kinases.* BMC Bioinformatics, 2013. 14: p. 345.

146. Whirl-Carrillo, M., et al., *Pharmacogenomics knowledge for personalized medicine.* Clin Pharmacol Ther, 2012. 92(4): p. 414-7.

147. Duran-Frigola, M., R. Mosca, and P. Aloy, *Structural systems pharmacology: the role of 3D structures in next-generation drug development.* Chem Biol, 2013. 20(5): p. 674-84.

148. Relling, M.V. and T.E. Klein, *CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network.* Clin Pharmacol Ther, 2011. 89(3): p. 464-7.

149. Landrum, M.J., et al., *ClinVar: public archive of relationships among sequence variation and human phenotype.* Nucleic Acids Res, 2014. 42(1): p. D980-5.

150. UniProt, C., *Reorganizing the protein space at the Universal Protein Resource (UniProt).* Nucleic Acids Res, 2012. 40(Database issue): p. D71-5.

151. NHLBI Exome Sequencing Project. *Exome Variant Server*. 2013 August 30 [cited 2012 December 12]; Available from: http://evs.gs.washington.edu/EVS/.

152. Law, V., et al., *DrugBank 4.0: shedding new light on drug metabolism.* Nucleic Acids Res, 2014. 42(Database issue): p. D1091-7.

153. McGuffin, L.J., M.T. Buenavista, and D.B. Roche, *The ModFOLD4 server for the quality assessment of 3D protein models.* Nucleic Acids Res, 2013. 41(Web Server issue): p. W368-72.

154. Dunbrack, R.L., Jr., *Rotamer libraries in the 21st century.* Curr Opin Struct Biol, 2002. 12(4): p. 431-40.

155. Hunter, S., et al., *InterPro in 2011: new developments in the family and domain prediction database.* Nucleic Acids Res, 2012. 40(Database issue): p. D306-12.

156. Pommie, C., et al., *IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties.* J Mol Recognit, 2004. 17(1): p. 17-32.

157. Malkov, S.N., et al., *A reexamination of the propensities of amino acids towards a particular secondary structure: classification of amino acids based on their chemical structure.* J Mol Model, 2008. 14(8): p. 769-75.

158. Rodriguez, R., et al., *Homology modeling, model and software evaluation: three related resources.* Bioinformatics, 1998. 14(6): p. 523-528.

159. Camps, J., et al., *FlexServ: an integrated tool for the analysis of protein flexibility.* Bioinformatics, 2009. 25(13): p. 1709-10.

160. Bosshard, H.R., D.N. Marti, and I. Jelesarov, *Protein stabilization by salt bridges: concepts, experimental approaches and clarification of some misunderstandings.* Journal of Molecular Recognition, 2004. 17(1): p. 1-16.

161. Petsko, G.A. and D. Ringe, *From sequence to structure*, in *Protein structure and function*, G.A. Petsko and D. Ringe, Editors. 2004, Sinauer Associates: Sunderland, MA.

162. Lahti, J.L., et al., *Bioinformatics and variability in drug response: a protein structural perspective.* J R Soc Interface, 2012. 9(72): p. 1409-37.

163. Zawaira, A., et al., *An expanded, unified substrate recognition site map for mammalian cytochrome P450s: analysis of molecular interactions between 15 mammalian CYP450 isoforms and 868 substrates.* Curr Drug Metab, 2011. 12(7): p. 684-700.

164. Ashkenazy, H., et al., *ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids.* Nucleic Acids Res, 2010. 38(Web Server issue): p. W529-33.

165. Rutherford, K. and V. Daggett, *Four human thiopurine s-methyltransferase alleles severely affect protein structure and dynamics.* J Mol Biol, 2008. 379(4): p. 803-14.

166. Sim, S.C. and M. Ingelman-Sundberg, *The Human Cytochrome P450 (CYP) Allele Nomenclature website: a peer-reviewed database of CYP variants and their associated effects.* Hum Genomics, 2010. 4(4): p. 278-81.

167. Rees, M.I., *ADVANCES IN PROTEIN CHEMISTRY AND STRUCTURAL BIOLOGY Challenges and Opportunities of Next-Generation Sequencing for Biomedical Research PREFACE.* Challenges and Opportunities of Next-Generation Sequencing for Biomedical Research, 2012. 89: p. Vii-Ix.

168. Berka, K., et al., *Membrane position of ibuprofen agrees with suggested access path entrance to cytochrome P450 2C9 active site.* J Phys Chem A, 2011. 115(41): p. 11248-55.

169. Cojocaru, V., et al., *Structure and Dynamics of the Membrane-Bound Cytochrome P450 2C9.* Plos Computational Biology, 2011. 7(8).

170. Wang, A., et al., *Crystal structure of human cytochrome P450 2D6 with prinomastat bound.* J Biol Chem, 2012. 287(14): p. 10834-43.

171. Genomes Project Consortium, et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. 491(7422): p. 56-65.

172. Lee, P.H. and H. Shatkay, *An integrative scoring system for ranking SNPs by their potential deleterious effects.* Bioinformatics, 2009. 25(8): p. 1048-55.

173. Biesecker, L.G., K.V. Shianna, and J.C. Mullikin, *Exome sequencing: the expert view.* Genome Biol, 2011. 12(9): p. 128.

174. Kiezun, A., et al., *Exome sequencing and the genetic basis of complex traits.* Nat Genet, 2012. 44(6): p. 623-30.

175. Gaedigk, A., *Complexities of CYP2D6 gene analysis and interpretation.* Int Rev Psychiatry, 2013. 25(5): p. 534-53.

176. Lahiry, P., et al., *Kinase mutations in human disease: interpreting genotype-phenotype relationships.* Nat Rev Genet, 2010. 11(1): p. 60-74.

177. Mullins, J.G.L., *Structural Modelling Pipelines in Next Generation Sequencing Projects.* Challenges and Opportunities of Next-Generation Sequencing for Biomedical Research, 2012. 89: p. 117-167.