ABSTRACT

Title of Dissertation:	COMPUTATIONAL METHODS IN MISSENSE MUTATION ANALYSIS: PHENOTYPES, PATHOGENICITY, AND PROTEIN ENGINEERING
	Yizhou Yin, Doctor of Philosophy, 2017
Dissertation directed by:	Professor John Moult Department of Cell Biology and Molecular Genetics

Understanding the molecular, phenotypic, and pathogenic effects of mutations is of enormous importance in human disease research and protein engineering. Both create a demand for computational methods to leverage the explosion of new sequence data and to explore the vast space of possible protein modifications and designs. My study in this dissertation demonstrates the value of computational methods in these areas. First, I developed a new ensemble method to predict continuous phenotype values as well as binary pathogenicity and objectively tested it in CAGI (Critical Assessment of Genome Interpretation). In two recent CAGI challenges, the method was ranked third in predicting the enzyme activity of missense mutations in NAGLU (N-Acetyl-Alpha-Glucosaminidase) and second in predicting the relative growth rate of mutated human SUMO-ligase in a yeast complementation assay. I also demonstrated the effectiveness of the new ensemble method for addressing a key problem limiting the use of current

mutation interpretation methods in the clinic – identifying which mutations can be assigned a pathogenic or benign status with high confidence. Next, I characterized and compared missense variants in monogenic disease and in cancer. The study revealed a number of properties of mutations in these two types of diseases, including: (a) methods based on sequence conservation properties are as effective for identifying cancer driver mutations as they are for monogenic disease mutations; (b) mutations in disordered regions of protein structure play a relatively small role in both classes of disease; (c) oncogenic mutations tend to be on the protein surface while tumor suppressors are concentrated in the core; (d) a large fraction of tumor suppressors act by destabilizing protein structure and (e) mutations in passenger genes display a surprisingly high level of deleteriousness. Finally, I applied computational methods to screen for appropriate mutations to enhance the thermostability of a catalytic domain of PlyC. This bacteriophage-derived endolysin has been demonstrated to have antimicrobial potential but its potential use is limited by its inherent thermosuseptibility. To prioritize stabilizing mutations, I conducted a rapid exhaustive survey of point mutations followed by validation using protein modeling and expert knowledge. The approach yielded three stabilizing mutants experimentally verified by our collaborators, with one particularly successful in terms of both thermal denaturation temperature and kinetic stability.

COMPUTATIONAL METHODS IN MISSENSE MUTATION ANALYSIS: PHENOTYPES, PATHOGENICITY, AND PROTEIN ENGINEERING

By

Yizhou Yin

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Professor John Moult, Chair
Professor Philip Bryan
Professor Eric Haag
Associate Professor Stephen Mount
Associate Professor Daniel Nelson

Foreword

I made substantial contributions to the relevant aspects of the jointly authored work included in the dissertation.

Chapter 2 was published in the *Human Mutation*:

Yin Y, Kundu K, Pal LR, Moult J. 2017. *Human Mutation*: 38(9): 1109-1122.

My contribution: computational experiments and data analysis

Chapter 4 was published in the *Protein Engineering, Design & Selection*:

Heselpoth RD, Yin Y, Moult J, Nelson DC. 2015. *Protein Engineering, Design & Selection*: 28(4):85-92.

My contribution: computational experiments and data analysis

Dedication

To Jiaqi, Jianping, and Lane

Acknowledgements

First and foremost, I would like to thank my advisor, John Moult, for his excellent mentorship during my graduate study. I didn't know much about bioinformatics and computational biology before I started in the Moult lab. John opened a door for me and guided me in this unfamiliar yet fascinating world with great wisdom and patience. He has always been supportive and understanding. More importantly, He showed me how to think and work as a real scientist, though I wish that I could have learned more from him. Those long and inspiring discussions with John will always be my precious memory.

I would like to thank the members of my committee along with the University of Maryland and IBBR faculty that offered me tremendous help. I'd like to thank Profession Daniel Nelson for the excellent collaboration work, which also helped make this dissertation possible. Deep gratitude to Professor Philip Bryan, Professor Stephen Mount, Professor Eric Haag for all their support. I'd like to thank Professor Osnat Herzberg, Professor Leslie Pick, Professor Edward Eisenstein, Professor Lindley Darden, Professor Ron Unger, Professor Arlin Stoltzfus, Professor Shunyuan Xiao, John Norvell, and many others that helped me through my graduate study.

My sincere gratitude to all my lab members, previous and current. Dr. Lipika Ray Pal, Dr. Zhen Shi, Dr. Chen-Hsin Yu, Dr. Chen Cao, Dr. Nuttinee Teerakulkittipong,

Kunal Kundu, and Maya Zuhl, as well as other fellow students in IBBR. They are more than just colleagues but also friends. My wholehearted thanks to the Woods family, who offered a warm home to an international student.

Lastly, I would like to thank my family, who were always understanding and encouraging with me during my graduate student time. They are the heroes in my life. Thanks to my wife, my most important companion in this long journey. And thanks to my parents, for their unconditional love.

Table of Contents

This Table of Contents is automatically generated by MS Word, linked to the Heading formats used within the Chapter text.

List of Tables

List of Figures

Chapter 3

Chapter 4

List of Abbreviations

In alphabetical order

ACMG	American College of Medical Genetics
APC	Adenomatosis Polyposis Coli, WNT signaling pathway regulator
AUC	Area under curve
BLCA	Bladder urothelial carcinoma
BRCA	Breast invasive carcinoma
CADD	Combined Annotation Dependent Depletion
CAGI	Critical Assessment of Genome Interpretation
CASP	Critical Assessment of protein Structure Prediction
CASR	Calcium sensing receptor
CBD	Cell wall binding domain
CD	Circular dichroism
CGC	Cancer Gene Census
CHAP	Cysteine, histidine-dependent amidohydrolase/peptidase
CHASM	Cancer-specific High-throughput Annotation of Somatic Mutations
ClinVar	NCBI database of genomic variation and its relationship to disease
CNV	Copy number variation
COADREAD	Colon and rectum adenocarcinoma
CONDEL	CONsensus DELeteriousness score of missense SNVs
COSMIC	Catalogue of Somatic Mutations in Cancer
CpG	5’-C-phosphate-G-3’ dinucleotide
CRAVAT	Cancer-Related Analysis of Variants Toolkit
dbNSFP	Database developed for functional prediction and annotation of all potential non-synonymous single-nucleotide variants (nsSNVs) in the human genome
DEE	Dead-end elimination
DISOPRED	Disordered residue predictor
DNA	Deoxyribonucleic acid
DSC	Differential scanning calorimetry
EAD	Enzymatically active domain
ExAC	The Exome Aggregation Consortium
FATHMM	Functional Analysis through Hidden Markov Models
FMPP	Familial Male-Limited Precocious Puberty
FOA	Fraction of agreement (among predictors)
FOLDX	A force field for energy calculations and protein design
G	Gibb’s free energy
GAS	Group A streptococcus
GBM	Glioblastoma multiforme
GCS	Group C streptococcus
GES	Group E streptococcus
GMEC	Global Minimum Energy Conformation
GTEx	Genotype-Tissue Expression project
GWAS	Genome-wide association study
GyH	Glycosyl hydrolase
HGMD	Human Gene Mutation Database

HNSC	Head and neck squamous cell carcinoma
HomoloGene	NCBI automated system for constructing putative homology groups from the complete gene sets of a wide range of eukaryotic species
IBBR	Institute for Bioscience and Biotechnology Research
ICGC	International Cancer Genome Consortium
IDR	Intrinsically disordered region
IMAC	Immobilized metal affinity chromatography
KIRC	Kidney renal clear-cell carcinoma
LAML	Acute myeloid leukemia
LB	Luria-Beterani
LHCGR	Luteinizing Hormone/Choriogonadotropin Receptor
LRT	The Likelihood ratio test
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MCMC	Markov Chain Monte Carlo
MIM	Mendelian Inheritance in Man
MRE	Mean residue ellipticity
NAGLU	N-Acetyl-Alpha-Glucosaminidase
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
NPV	Negative predictive value
NRMSD	Normalized root mean square deviation
OMIM	Online Mendelian Inheritance in Man
OD	Optical density
OV	Ovarian serous cystadenocarcinoma
PAH	Phenylalanine hydroxylase
PBS	Phosphate buffered saline
PCAWG	PanCancer Analysis of Whole Genomes
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PDF	Predicted deleterious/destabilizing fraction
PONP	The Pathogenic-or-Not-Pipeline
PPH2	Polyphen-2
PPV	Positive predictive value
PROVEAN	Protein Variation Effect Analyzer
RANGAP1	Ran GTPase Activating Protein 1
RBF	Radial Basis Function kernel
REVEL	Rare Exome Variant Ensemble Learner
RMSD	Root mean square deviation
ROC	Receiver operating characteristic
ROSETTA	A software suite for structure prediction and protein design
RPM	Rounds per minute
SASA	Solvent Accessible Surface Area
SCWRL	Side-Chains With a Rotamer Library
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEM	Standard error of the mean

SIFT	The Sorting Intolerant from Tolerant algorithm
SNAP2	Screening for non-acceptable polymorphisms
SNP	Single-nucleotide polymorphism
SNPs3D	Predictor assigning molecular functional effect of non-synonymous SNPs based on structure and sequence analysis
SNV	Single-nucleotide variation
STRIDE	Structural Identification, algorithm for the assignment of protein secondary structure elements
SUMO	Small Ubiquitin-like Modifier
SVM	Support vector machine
$T_{1/2}$	Circular dichroism melt equilibrium temperature
T_G	Thermal transition temperature
T_m	Melting temperature
TCGA	The Cancer Genome Atlas
TSG	Tumor suppressor gene
UBE2I	Ubiquitin Conjugating Enzyme E2 I
UBI4	Ubiquitin
UCEC	Uterine corpus endometrioid carcinoma
UniProt	Database of protein sequence and functional information
USPTO	United States Patent and Trademark Office
UV	Ultraviolet
VEST3	Variant Effect Scoring Tool
Weka	Waikato Environment for Knowledge Analysis, software suite for machine learning
WT	Wild-type

Chapter 1: Introduction

1.1 Missense variants and human diseases

1.1.1 The landscape of human mutations

DNA mutations spontaneously occur at very low frequency due to replication, repair and mitotic error. They can also result from exogenous and endogenous factors such as chemicals, ultraviolet light, ionizing radiation, oxygen free radicals, and viruses. As a consequence, mutations inevitably accumulate in both germ cells and somatic cells in spite of the high fidelity molecular machinery for replicating and repairing DNA. There are many types of genetic variations including SNVs (single nucleotide variants), indels (insertions and/or deletions), CNVs (copy number variations), chromosome rearrangements, and large-scale events (e.g. aneuploidy, chromothripsis). In terms of the protein translation, SNVs can be synonymous (no amino acid change), or nonsynonymous, including missense that alters amino acids, and nonsense that leads to premature termination. Non-coding SNVs can affect splicing, alter expression and other regulatory processes, or locate at intergenic regions with uncertain roles. The vast majority (96%) of genetic variations observed in populations are SNVs, although the very few structure variations (large indels, CNVs, chromosome rearrangements) affect more base pairs (1000 Genomes Project Consortium et al., 2015).

Advances in next-generation sequencing technologies (Reuter, Spacek, & Snyder, 2015; Shendure & Ji, 2008; Soon, Hariharan, & Snyder, 2013) have generated sequence data for tens of thousands of genomes and exomes and that has led to a deepening of our understanding of the landscape of human mutations. For germline mutations (mutations that occur in the germ cells), it is estimated that each individual human carries 1~5 million genetic variants in the whole genome compared with the reference genome (1000 Genomes Project Consortium et al., 2015; Roach et al., 2010), out of which 30~70 are *de novo* point mutations compared with the parents (Francioli et al., 2015). Although on average only about 0.3% of SNPs (single nucleotide polymorphisms) observed in one genome are nonsynonymous (and 0.3% synonymous, 13% regulatory, and 47% in intron), there are ~10,000 nonsynonymous SNVs in each genome when compared with the reference genome (1000 Genomes Project Consortium et al., 2015; Ng et al., 2008). On top of these, through trillions of cell divisions from early development to adulthood, a substantial number of somatic mutations (mutations that occur in the somatic cells) accumulate over time. For instance, by middle age, thousands of point mutations may have accumulated in the sun-exposed skin cells (Martincorena et al., 2015). As another example, it is roughly estimated that there are totally a billion independent mutations accumulated in the whole intestinal epithelium of a 60-year-old individual (Lynch, 2010). Cancer cells in an individual typically carry various numbers (1000 ~20,000) of somatic point mutations, out of which from 10 to as many as 1000 are nonsynonymous (Vogelstein et al., 2013).

The mutation rates of germline SNVs has been estimated as between ~1.0×10^{-8} to 2.2×10^{-8} per base pair (bp) per generation, varying depending on the approach (high penetrant Mendelian disease (Lynch, 2010), Phylogenetic analysis (Chimpanzee Sequencing and Analysis Consortium, 2005), or whole genome sequencing of pedigrees (Ségurel, Wyman, & Przeworski, 2014). It has also been reported to vary by more than 100-fold within the genome across individuals (Michaelson et al., 2012). Germline SNVs tend to be enriched in certain sequence compositions, especially at CpG dinucleotides (Hwang & Green, 2004). The mutation rates of germline mutations in coding regions are strongly constrained by the expression level of genes (Drummond, Raval, & Wilke, 2006; Drummond & Wilke, 2008). The mutation rate is also affected by sex and parental age in that *de novo* point mutations among offspring are predominantly related to paternal age (Kong et al., 2012) and chromosomal nondisjunction errors are mainly affected by maternal age (Sherman et al., 1994). Fewer studies have estimated the mutation rates for other types of variants partially because it is technically challenging (Shendure & Akey, 2015). As an example, it is estimated that 2.94 small indels ($\leq$ 20 bp) and 0.16 structural variants (> 20 bp) occur per generation (Kloosterman et al., 2015). The estimated mutation rate of somatic mutations is around one order of magnitude higher than germline mutations (Lynch, 2010). It dramatically varies across cancer types and individuals by up to two orders of magnitude (Martincorena & Campbell, 2015; Vogelstein et al., 2013). On a fine scale, the mutation rates in somatic cells vary depending on environmental factors and impaired DNA replication or repair (Shendure & Akey, 2015). On a chromosomal scale, the variation is largely determined by chromatin

organization (Martincorena & Luscombe, 2013; Schuster-Böckler & Lehner, 2012; Shendure & Akey, 2015).

1.1.2 Human diseases and their genetic basis

There are three major types of human diseases that closely link to genetic variations, monogenic disease (or Mendelian disease), complex trait disease, and cancer. The monogenic diseases are usually caused by mutations in a single gene or one of a few disease genes. They can follow either a dominant or recessive inheritance pattern and are mostly rare. For example, one of the lysosomal storage diseases, Sanfilippo syndrome IIIB is caused by mutations in *NAGLU* gene. The disease is autosomal recessive with a reported birth incidence of 0.28-4.1 per 100,000 (Valstar, Ruijter, van Diggelen, Poorthuis, & Wijburg, 2008). To date, more than 7000 monogenic diseases have been catalogued in the Online Mendelian Inheritance in Man (OMIM) database (http://omim.org/). Despite the very low incidences of individual monogenic diseases, the birth prevalence of all monogenic diseases in industrialized countries was estimated to be 3.6 per 1,000 newborns (Baird, Anderson, Newcombe, & Lowry, 1988). This number is even higher in the developing countries. Disease-causing genes have been identified for more than half of the rare monogenic diseases (Boycott, Vanstone, Bulman, & MacKenzie, 2013). The Human Gene Mutation Database (HGMD) is a major database of monogenic disease-related genes and mutations. Currently, it includes over 203,000 unique gene lesions in over 8000 genes for inheritable disease collected from literature (Stenson et al., 2017). 56% of these lesions are missense or nonsense SNVs. In an early version that contains a much

higher fraction of rare monogenic diseases, this proportion is about 60% (Stenson et al., 2003).

Complex trait diseases have a higher rate of occurrence than monogenic diseases. A single complex trait disease can have incidence and prevalence similar to or more than all monogenic diseases combined. For example, Crohn's disease affects about two million people in North America (Molodecky et al., 2012). The number is 5.3 million for Alzheimer's disease (Alzheimer's Association, 2015). Globally, diabetes may affect 439 million adults by 2030 (Shaw, Sicree, & Zimmet, 2010). Unlike monogenic diseases, up to hundreds of loci in the genome may contribute to a single complex trait disease (de Lange et al., 2017). Variants in many of these loci only make a small contribution to a disease phenotype. Complex trait diseases are also heavily affected by environmental and behavioral factors. For many, the relevant genes and variants are still not clear. Much information has been obtained by the genome-wide association studies (GWAS), which captures disease-associated common SNPs using microarray technology. By including several thousand individuals with and without diseases, more than 50,000 unique SNP-trait associations have been discovered in more than 2,500 studies (GWAS catalog, https://www.ebi.ac.uk/gwas/). Through these associated loci, disease-causing mechanism SNPs may sometimes be imputed. It has been shown that missense mutations also play an important role in complex trait diseases (Kryukov, Pennacchio, & Sunyaev, 2007; Pal & Moult, 2015).

Cancers are diseases where cells escape constraints on growth, and potentially invade other parts of the body. Most cancers are caused by somatic mutations, while certain germline mutations can increase the risk of an individual developing the disease. There have been sequencing studies at the level of exomes and increasingly complete in more than 30 types of cancers. The sequence data are available in several major databases including the Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/), Catalogue of Somatic Mutations in Cancer (COSMIC, http://cancer.sanger.ac.uk/cosmic), and the International Cancer Genome Consortium (ICGC, http://icgc.org). To date, more than 500 genes have been catalogued by the Cancer Gene Census (CGC, http://cancer.sanger.ac.uk/census) as causal genes. Although dramatic variations exist across individuals and across cancer types, there are on average 33 to 66 mutated genes with altered functions in a tumor. 86% of the mutations in these genes are missense, 7% are nonsense, and 1.6% are at splice sites or close to coding regions (Vogelstein et al., 2013). It was also reported that cancer types can be divided into two classes based on whether dominated by SNVs or by CNVs (Ciriello et al., 2013). At one end of the spectrum, a single tumor can carry thousands of mutations if the mismatch repair (Gryfe & Gallinger, 2001) or proofreading machinery (Palles et al., 2013) is damaged. At the other end of the spectrum, on average pediatric tumors and leukemias may just carry 9.6 point mutations per tumor (Vogelstein et al., 2013). In contrast to this large variation, it is estimated that most cancer types have less than 5 key point mutations (real driver) per tumor (Sabarinathan et al., 2017). At present, there are no established quantitative models on the origins of driver mutations. It has been shown that risk of different

cancer types is closely related to the number of cell divisions in the corresponding tissue (Tomasetti, Li, & Vogelstein, 2017), implying that driver incidence is largely determined by this factor. In general, there are two types of cancer driver genes, oncogenes that acquire gain-of-functions through mutations, and tumor suppressor genes that lose function through mutations. Driver mutations in well-studied oncogenes and tumor suppressor genes show obvious different patterns, with mutations recurrently happening at the same positions (hotspots) in oncogenes, and more evenly distributed through tumor suppressor genes (Vogelstein et al., 2013). One feature of cancer mutations is that mutations in cancer driver genes can also be passenger mutations. For example, among mutations in the APC protein, only those within the N-terminus are drivers, whereas those within the C-terminus are passengers (Vogelstein et al., 2013).

1.2 Computational interpretation of missense mutations

1.2.1 General mutation interpreting methods

The advent of massive genome and exome sequencing creates a major demand for reliable bioinformatics tools to interpret and prioritize the genetic variations that have functional consequences. New computational approaches have been developed and applied to genetic variations in general (Cooper & Shendure, 2011; Peterson, Doughty, & Kann, 2013) and in cancer (Gonzalez-Perez, Mustonen, et al., 2013). In principle, a mutation interpretation method can identify the consequence of a given genetic variant at various levels: 1) functional impact at the molecular level, 2)

deleteriousness to the organism, and 3) pathogenicity (whether or not causing disease) (Shendure & Akey, 2015). While these three tasks are related, the majority of current computational methods measure deleteriousness using evolution information. For missense mutations, methods compare a mutation substitution with residues found in homologous protein sequences and variants within the human population under the assumption that conserved positions or the absence of population variants indicate stronger constraints from purifying selection. In this sense, these methods make use of fitness impact as a surrogate for pathogenicity (Calabrese, Capriotti, Fariselli, Martelli, & Casadio, 2009; Choi, Sims, Murphy, Miller, & Chan, 2012; Chun & Fay, 2009; Katsonis & Lichtarge, 2014; Kircher et al., 2014; Lichtarge, Bourne, & Cohen, 1996; Ng & Henikoff, 2003; Niroula & Vihinen, 2016; Schwarz, Rödelsperger, Schuelke, & Seelow, 2010; Thomas et al., 2006; Yue & Moult, 2006). The similar principle can be applied to methods that address non-coding variations, where nucleotide sequence profiles replace protein sequence profiles (Cooper et al., 2005; Pollard, Hubisz, Rosenbloom, & Siepel, 2010). Some methods also incorporate physical-chemical (e.g. amino acid properties), structure information (e.g. secondary structure element and solvent accessibility), and functional annotations (Adzhubei et al., 2010; Baugh et al., 2016; Carter, Douville, Stenson, Cooper, & Karchin, 2013; Folkman, Stantic, Sattar, & Zhou, 2016; B. Li et al., 2009). A few methods adopt an ensemble approach by incorporating outcomes from multiple other methods (Capriotti, Altman, & Bromberg, 2013; González-Pérez & López-Bigas, 2011; Ioannidis et al., 2016; Olatubosun, Väliaho, Härkönen, Thusberg, & Vihinen, 2012). A few methods seek a different approach by integrating three-dimensional structure

modeling to infer impact on protein thermostability (Redler, Das, Diaz, & Dokholyan, 2016; Yue, Li, & Moult, 2005). These methods are designed to detect when a mutation destabilizes protein three-dimensional structure, and so have more limited scope than sequence methods. Most methods use supervised machine learning and require a training classifier such as random forest (Carter et al., 2013; B. Li et al., 2009; Niroula, Urolagin, & Vihinen, 2015), neural network (Hecht, Bromberg, & Rost, 2015), or support vector machines (SVMs) (Calabrese et al., 2009; Kircher et al., 2014; Yue & Moult, 2006). A few methods rely on direct measures of certain properties (e.g. evolutionary), and do not require training (Choi et al., 2012; Chun & Fay, 2009; Lichtarge et al., 1996; Ng & Henikoff, 2003; Thomas et al., 2006).

1.2.2 Cancer-specific methods

Although originally not intended for that purpose, most methods mentioned above can be applied to interpret the impact of cancer somatic mutations, as will be discussed later in Chapter 3. In addition, there are computational methods specifically developed for interpreting cancer data. One class of methods aims to prioritize cancer driver genes using mainly three types of information (Hofree et al., 2016): 1) SNV recurrence, 2) SNV molecular impact, and 3) SNV spatial clustering. Some only rely on SNV molecular impact (Gonzalez-Perez, Deu-Pons, & Lopez-Bigas, 2012) or SNV clustering (Tamborero, Gonzalez-Perez, Perez-Llamas, et al., 2013; Tamborero, Gonzalez-Perez, & Lopez-Bigas, 2013). Others combined all three types of information (Dees et al., 2012; Khurana et al., 2013; Lawrence et al., 2013). The second class of methods aims to prioritize cancer somatic mutations using similar

information to that in the general mutation interpreting methods (Carter et al., 2009; Gonzalez-Perez et al., 2012; Joshua S Kaminker, Zhang, Watanabe, & Zhang, 2007; Mao et al., 2013; Reva, Antipin, & Sander, 2011; Shihab, Gough, Cooper, Day, & Gaunt, 2013; Yue et al., 2010). Very few computational methods were designed to address other types of somatic variations (e.g. CNV) (Mermel et al., 2011), or mutated gene subnetworks (Leiserson et al., 2015). There is also one method that combines driver gene discovery and mutation analysis into a single pipeline (Gonzalez-Perez, Perez-Llamas, et al., 2013).

1.2.3 Critical assessment of contemporary methods

There are very few studies that independently assess the performance of the current mutation interpretation methods (Gnad, Baucom, Mukhyala, Manning, & Zhang, 2013; Martelotto et al., 2014). In order to have an objective assessment of the state of the art, John Moult and Steven Brenner started the Critical Assessment of Genome Interpretation (CAGI, https://genomeinterpretation.org/) (Hoskins et al., 2017) as community-wide experiments to test a variety of mutation and genome interpretation methods. In analogy to the Critical Assessment of Structure Prediction (CASP) (Moult, Fidelis, Kryshtafovych, Schwede, & Tramontano, 2016), CAGI strictly separates predictors, data providers, and assessors. Participants are asked to predict particular phenotypes, given genetic variant information. Meanwhile, the corresponding experimental results are not released until all participants have submitted their predictions, thus these are *bona fide* blind predictions. Independent experts assess the predictions and the outcomes are discussed at a CAGI conference.

The challenges in CAGI cover a wide range of prediction problems and datasets. For example, in the latest CAGI round (CAGI4), there were datasets of germline and somatic mutations in exomes, whole genomes, clinical gene panels and individual genes in the context of rare monogenic disease, complex trait disease, and cancers. Challenges also included identification of eQTL causal SNPs and deep mutation scanning data. Most missense mutation analysis methods report a binary assignment of deleterious or not deleterious. Therefore, two CAGI challenges, NAGLU and SUMO-ligase (Zhang et al., 2017), are of particular interest in this dissertation in that they request predictions of continuous activity values. Initially motivated by this, I developed an ensemble approach, and further extended its application to both binary pathogenic prediction and estimation of the subset of mutations with high-reliability assignments.

The human *NAGLU* gene encodes N-acetyl-glucosaminidase, an enzyme that catalyzes the cleavage of the glucosaminoglycan chain of heparin sulfate in lysosomes. It is one of four genes (Valstar et al., 2008) in which mutations may cause one of the four types of Mucopolysaccharidosis III or Sanfilippo Syndrome (Sanfilippo, Podosin, Langer, & Good, 1963), a severe neurological disease. The human NAGLU protein exists as a homo-trimer *in vivo*. The three-dimensional structure of the protein is available from a recent patent (US08775146B2, 2014) (Figure 1-1). The CAGI challenge was based on a set of 165 rare missense NAGLU mutations found in the European population samples of the ExAC exome database (Lek et al., 2016) (excluding known disease-causing mutations). BioMarin, the

company providing the challenge data, introduced each mutation into a human cell line via a plasmid construct, and after a period of cell growth, measured NAGLU enzyme activity in the cell lysate.

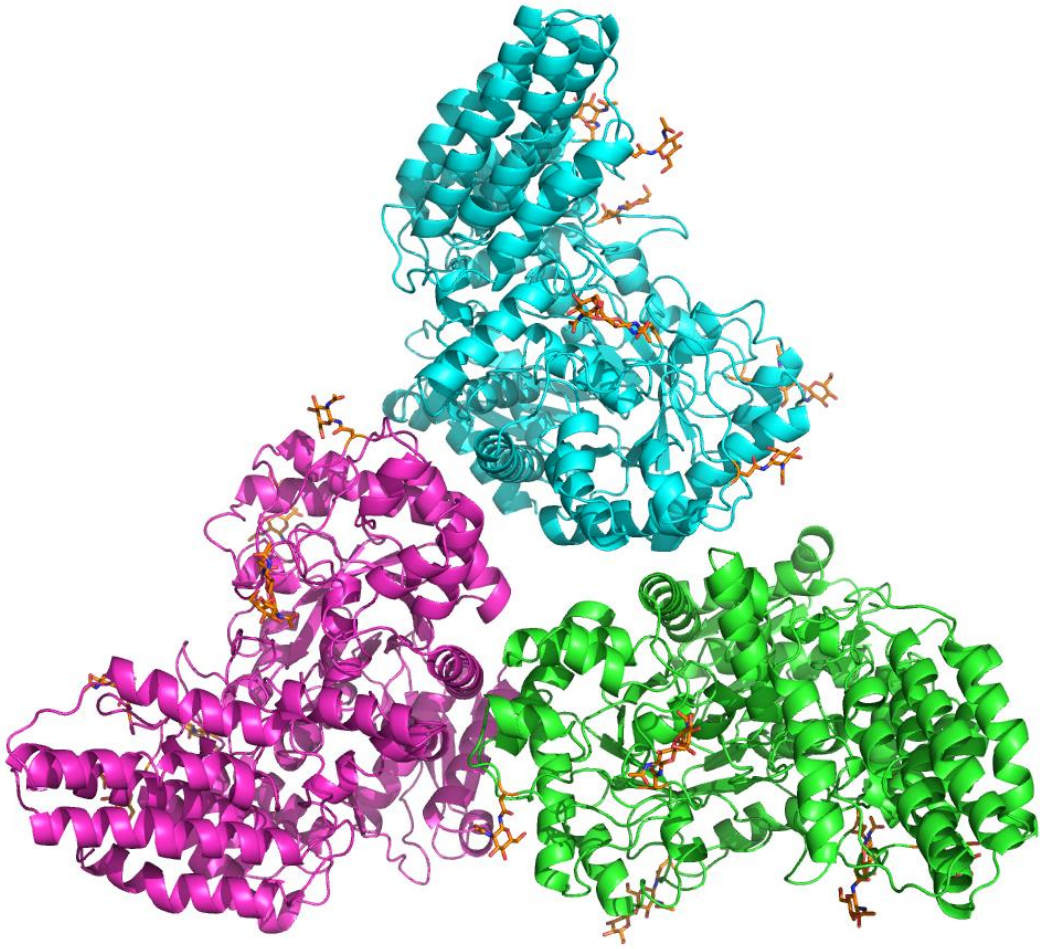


Figure 1-1. Structure of the human NAGLU homo-trimer, based on the crystal structure reported in patent USPTO US08775146B2. The three identical NAGLU monomers (green, cyan, and magenta) form a symmetrical complex. The glycosylation sites and the glycan molecules are shown in orange.

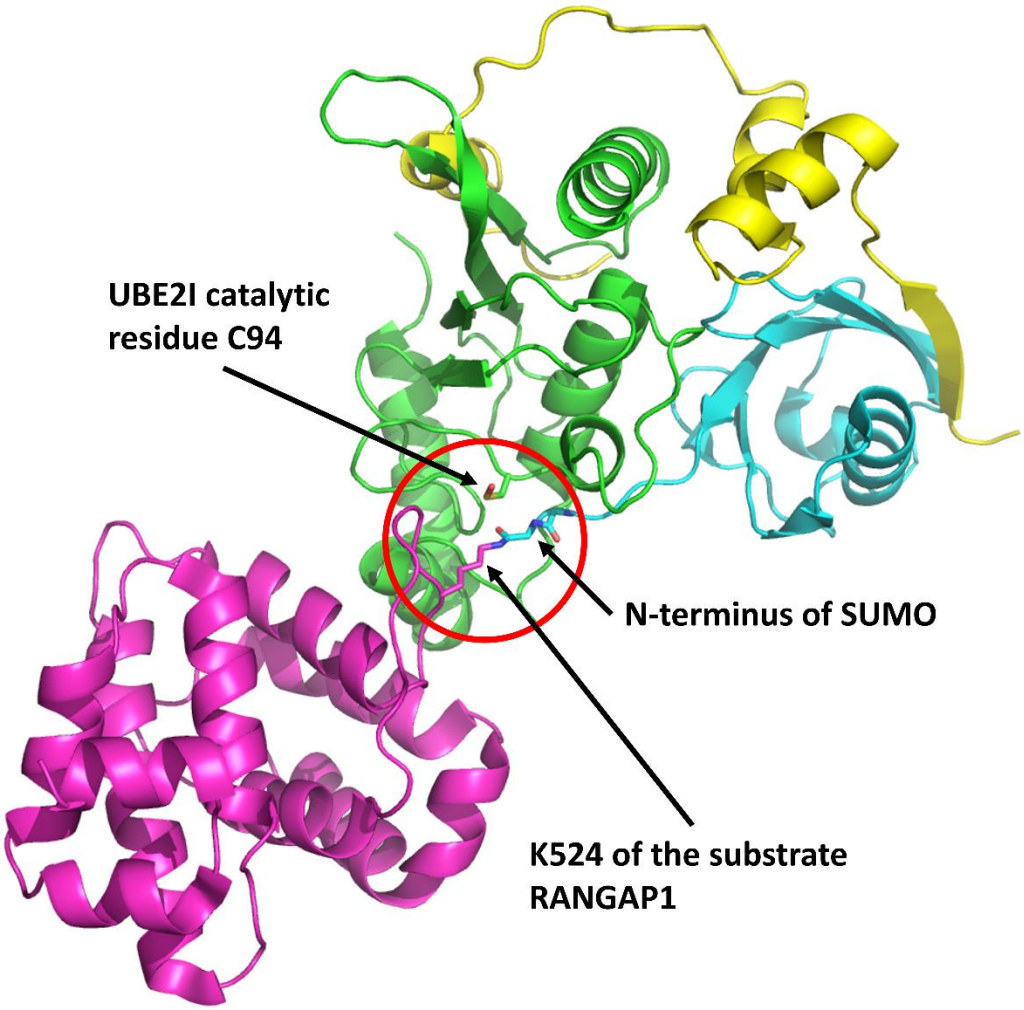


Figure 1-2. Structure of human UBE2I (green) in complex with SUMO (cyan), E3 ligase (yellow), and the substrate RANGAP1 (magenta). The UBE2I catalytic site is indicated by the red circle. PDB: 3UIP (Gareau, Reverter, & Lima, 2012)

The human *UBE2I* gene encodes the SUMO E2 ligase, an enzyme that catalyzes the attachment of a range of target substrate proteins to SUMO (Geiss-Friedlander & Melchior, 2007) (See Figure 1-2 for one example). In the challenge, over 6,000 human SUMO-ligase mutant genes were tagged with DNA barcodes and cloned into

S. cerevisiae cells carrying a temperature-sensitive *UBC9* gene (encoding the corresponding yeast SUMO-ligase (Weile et al., 2017). The relative abundance of cells carrying each mutant gene was deduced from high-throughput sequencing of the barcodes, following 48 hours of cell growth.

1.3 Engineering through protein design

1.3.1 Advances in structure modeling methods

The goal of computational protein design is to find a protein sequence that folds into an appropriate three-dimensional structure and so confers a desired new property or function such as a *de novo* fold, enhanced protein thermostability, altered binding affinity or specificity of protein-ligand and protein-macromolecule interaction, altered enzymatic activity, among others. Protein design methods typically involve a process in which an energy function and a conformational search process are used to assess the impact of some particular missense mutations in a certain structure context, which is, in some sense, comparable to the structure based mutation interpreting tools. However, unlike the latter, protein design also involves a process to search for optimal mutations in a large sequence space. In full protein design and in mutation selection, experimental validation plays a key role.

Protein design methods arose out of protein structural modeling methods, for example, the RosettaDesign protocol of the Rosetta programs (B Kuhlman & Baker, 2000; Rohl, Strauss, Misura, & Baker, 2004). In the past two decades, tremendous

improvements have been made to computational protein modeling algorithms (Samish, 2017). A few examples: the use of a set of discrete rotamers and conformers instead of treating amino acid side chain conformations in a continuous three-dimensional space facilitates the efficacy of describing protein structures (Dunbrack, 2002). The inclusion of flexible protein backbone sampling methods improves the accuracy of the structure modeling (Ollikainen, Smith, Fraser, & Kortemme, 2013). The invention of a knowledge-based approach to reassemble proteins from known high-resolution structural fragments has greatly boosted the efficiency and accuracy of protein structure prediction (B Kuhlman & Baker, 2000). To search for the global minimum energy conformation (GMEC), both deterministic methods (e.g. linear programming, and dead-end elimination, DEE) and stochastic methods (e.g. Monte Carlo Markov Chain, MCMC) have been adopted (Samish, 2009). For protein design, improvements were also seen in terms of the design strategy, such as the inclusion of negative design and the Cluster Expansion method (Grigoryan, Reinke, & Keating, 2009). Protein design is a relatively new field, and there are still many challenges, such as loop design.

1.3.2 Application of computational protein design methods

Computational protein design has been applied to a broad range of protein engineering problems. For example, it was used to improve the thermostability of human acetylcholinesterase by 20°C (Goldenzweig et al., 2016) and increased the midpoints of thermal denaturation of Drosophila melanogaster homeodomain from

49°C to 99°C (Shah et al., 2007), to alter protein-protein binding specificity (Grigoryan et al., 2009), to design protein based lysozyme inhibitor (Procko et al., 2013), to design pH-dependent protein binding specificity (Strauch, Fleishman, & Baker, 2014), and to improve antibody affinity (Lippow, Wittrup, & Tidor, 2007). More prominently, computation protein design methods have successfully designed proteins with novel folds that do not exist in nature (Brian Kuhlman et al., 2003; Xiong et al., 2014), a small stable vaccine to induce potent neutralizing antibodies (Correia et al., 2014), and enzymes with new activity (Jiang et al., 2008; Röthlisberger et al., 2008; Siegel et al., 2010).

1.3.3 Rosetta and FoldX

The Rosetta program (Das & Baker, 2008) was developed by David Baker and many collaborators. The method consists of two major algorithms (B Kuhlman & Baker, 2000; Rohl et al., 2004): 1) Monte Carlo sampling of peptide fragment conformation space shaped by local contacts to reduce the conformational search problem 2) Searching for the optimal protein conformation using an energy function of mixed physical and statistical terms. It now includes dozens of modeling and design protocols tailored to different tasks. Particularly, the Rosetta ddG application predicts the change of protein folding free energy change ($\Delta\Delta G$) induced by point mutations (Kellogg, Leaver-Fay, & Baker, 2011). Another program, FoldX (Guerois, Nielsen, & Serrano, 2002; Schymkowitz, Borg, et al., 2005; Schymkowitz, Rousseau, et al., 2005) adopted an empirical energy function based on both physical terms such as van

der Waals interactions, hydrogen bonding, electrostatic and solvation, and statistical energy terms computed from observations in the database of known protein structures. Given a point mutation, FoldX uses a fixed backbone modeling process that can rapidly estimate the impact of a mutation on the folding free energy.

1.3.4 PlyC as a potential antimicrobial

Endolysins (phage lysins) are bacteriophage peptidoglycan hydrolases (Nelson, Schuch, Chahales, Zhu, & Fischetti, 2006). Purified PlyC proteins can be applied to lyse the cell wall of susceptible streptococci both *in vitro* and in *vivo* with a superior activity compared to other endolysins (Nelson, Loomis, & Fischetti, 2001). The protein structure of PlyC is unique in that it consists of eight identical PlyCB monomers forming an octameric ring structure as the cell wall binding domain, and a PlyCA subunit as the catalytic domain. Two domains of PlyCA, N-terminal glycosyl hydrolase (GyH) and C-terminal cysteine, histidine-dependent amidohydrolase/peptidase (CHAP), catalyze distinct reactions synergistically. The PlyCB octamer and PlyCA are connected via a helical docking domain (McGowan et al., 2012). Although PlyC is a potential antimicrobial, the PlyCA CHAP domain is very thermal unstable, which could limit its shelf life.

1.4 Overview

The dissertation is organized as follows. In Chapter 2, I first review the previous efforts and analyses in interpreting missense variants in human diseases, followed by

an introduction of the two CAGI challenges. I then present the results and post analysis of the ensemble methods for prediction of continuous activity, binary assignment, and estimation of reliability. In Chapter 3, I describe the current picture of missense somatic mutations in cancers, the current computational methods, and issues in identifying cancer driver genes and driver mutations. I then present the results of characterizing and comparing missense variants in monogenic diseases and cancer in light of a better understanding of cancer driver mutations and their effects. In Chapter 4, I describe the computational and experimental approaches to engineer PlyC CHAP domain thermostability. I then present the successful mutation and the experimental validation results. In Chapter 5, I summarize the conclusions of the three projects and then look into future prospects for improving computational mutation interpretation and protein engineering methods.

Chapter 2: Ensemble variant interpretation methods to predict enzyme activity and assign pathogenicity in the CAGI4 NAGLU (Human N-acetyl-glucosaminidase) and UBE2I (Human SUMO-ligase) challenges

Published:

Yin Y, Kundu K, Pal LR, Moult J. 2017. Ensemble variant interpretation methods to predict enzyme activity and assign pathogenicity in the CAGI4 NAGLU (Human N-acetyl-glucosaminidase) and UBE2I (Human SUMO-ligase) challenges. *Human Mutation* 38(9):1109-1122.

My contribution: computational experiments and data analysis

2.1 Abstract

CAGI (Critical Assessment of Genome Interpretation) conducts community experiments to determine the state of the art in relating genotype to phenotype. Here we report results obtained using newly-developed ensemble methods to address two CAGI4 challenges: enzyme activity for population missense variants found in *NAGLU* (Human N-acetyl-glucosaminidase) and random missense mutations in Human *UBE2I* (Human SUMO E2 ligase), assayed in a high throughput competitive yeast complementation procedure. The ensemble methods are effective, ranked 2nd for SUMO-ligase and 3rd for NAGLU, according to the CAGI independent assessors.

However, in common with other methods used in CAGI, there are large discrepancies between predicted and experimental activities for a subset of variants. Analysis of the structural context provides some insight into these. Post-challenge analysis shows the ensemble methods are also effective at assigning pathogenicity for the *NAGLU* variants. In the clinic, providing an estimate of the reliability of pathogenic assignments is key. I have also used the NAGLU dataset to show that ensemble methods have considerable potential for this task, and are already reliable enough for use with a subset of mutations.

2.2 Introduction

The vast quantities of data generated by the high-throughput genotyping and next-generation sequencing technologies (Reuter et al., 2015; Soon et al., 2013) have created a major demand for reliable methods of interpreting the phenotypic significance of genetic variation, particularly as it relates to human disease. Among various types of genetic variation, missense single nucleotide polymorphisms (SNPs) and missense rare mutations in coding regions are of particular interest because of the major role these play in monogenic disease (Stenson et al., 2014), complex trait disease (Kryukov et al., 2007; Pal & Moult, 2015), and cancer (Shi & Moult, 2011; Wood et al., 2007).

Many computational methods have been developed to identify the relevance of missense variants to disease (Peterson et al., 2013). Most of these methods make use of sequence variation across species and within the human population to infer the

likely fitness impact of an amino acid substitution, assumed to be related to disease relevance (Calabrese et al., 2009; Choi et al., 2012; Chun & Fay, 2009; Katsonis & Lichtarge, 2014; Kircher et al., 2014; Lichtarge et al., 1996; Ng & Henikoff, 2003; Niroula et al., 2015; Schwarz et al., 2010; Thomas et al., 2006; Yue & Moult, 2006). A few make use of three-dimensional structure information, particularly to infer any thermodynamic destabilization of the structure (Redler et al., 2016; Yue et al., 2005), assuming that decreased protein activity implies a relationship to disease. Some methods combine both sequence and structure information (Adzhubei et al., 2010; Baugh et al., 2016; Carter et al., 2013; Folkman et al., 2016; Hecht et al., 2015; B. Li et al., 2009). Methods usually use supervised machine learning such as random forest (Carter et al., 2013; B. Li et al., 2009; Niroula et al., 2015), neural network (Hecht et al., 2015) and support vector machines (Calabrese et al., 2009; Kircher et al., 2014; Yue & Moult, 2006), or models that do not need training (Choi et al., 2012; Chun & Fay, 2009; Lichtarge et al., 1996; Ng & Henikoff, 2003; Thomas et al., 2006).

Missense analysis methods have usually been evaluated by benchmarking against databases of known monogenic disease mutations and presumed benign species or population variants, and there have been very few independent tests. Critical Assessment of Genome Interpretation (CAGI), conducts community-wide experiments to test these and other genome interpretation methods. CAGI participants are provided genetic variant information and asked to predict phenotypic consequences. Independent assessors then evaluate the results. The experiments are double blind in that participants do not know the phenotypes and the assessors do not

know the identity of the participants. In the most recent CAGI round, CAGI4 (http://genomeinterpretation.org), there were two missense variant interpretation challenges: the NAGLU challenge (https://genomeinterpretation.org/content/4-NAGLU) and the SUMO-ligase challenge (https://genomeinterpretation.org/content/4-SUMO_ligase). Here we report our results for these.

NAGLU (MIM# 609701) encodes Human N-acetyl-glucosaminidase, an enzyme involved in the heparan sulfate degradation process, and is one of four (Valstar et al. 2008) lysosomal enzymes in which mutations may result in one of four corresponding types of Sanfilippo Syndrome (Sanfilippo et al. 1963). Mutations in NAGLU protein cause a rare neurological disease, Mucopolysaccharidosis IIIB or Sanfilippo B disease (O'Brien 1972; von Figura and Kresse 1972; Valstar et al. 2008). The NAGLU challenge utilized *in vitro* enzyme activity data for a set of 165 rare population missense mutations extracted from the ExAC exome database (60,706 individual genomes) (Lek et al., 2016), omitting 24 known disease mutations. CAGI challenge participants were asked to quantitatively predict the enzymatic activity of each mutant relative to that of the wild-type enzyme. A unique feature of the NAGLU dataset is that it represents the distribution of protein function of rare variants present in a population. To our knowledge, this is the first test of this type for current missense analysis methods, and more relevant to variants encountered in the clinic than usual database benchmarking.

UBE2I (MIM# 601661) encodes the human small ubiquitin-like modifier proteins conjugating protein (SUMO E2 ligase) that catalyzes the covalent attachment of SUMO to a range of target proteins. The CAGI challenge data provider had generated a library of over 6,000 human SUMO-ligase *UBE2I* clones expressing nearly 2,000 unique missense mutations in various combinations. The competitive growth rate of each clone was deduced from deep sequencing of a yeast-based complementation system. CAGI participants were asked to predict the relative competitive growth rates of yeast cells carrying three different sets of random mutations. Unlike the NAGLU challenge, where enzyme activity is known to be directly related to pathogenicity (von Figura & Kresse, 1972), the relationship between SUMO-ligase function and fitness is complicated by two factors – the multiple regulator and target proteins that interact with SUMO-ligase (Geiss-Friedlander & Melchior, 2007), and the fact that the human SUMO-ligase was substituted for the native enzyme in yeast cells. These factors make this a complex system from the point of view of interpreting the CAGI results. Many similar high throughput mutational scans are now being undertaken, so it is of interest to use the CAGI experiment to begin to probe the strengths and limitations of this approach, both generally, and as a basis for CAGI challenges.

All submitted predictions in each challenge were evaluated by independent assessors, one for each challenge. Results reported here were ranked 2nd among 9 groups with 16 submissions for the SUMO-ligase challenge and 3rd among 10 groups with 17 submissions for NAGLU.

Most missense analysis methods assign each variant as either deleterious or benign. An unusual feature of both the NAGLU and SUMO-ligase challenges is that they require prediction of a continuous variable, in one case relative enzyme activity, and in the other, relative yeast growth rate. In other words, the challenges require a regression predictor rather than a classification predictor. To address this requirement, I made use of an ensemble approach, combining binary predictions or associated confidence scores from up to eleven different methods. In a number of fields, ensemble methods that combine results from multiple individual methods have proven effective (Abeel, Helleputte, Van de Peer, Dupont, & Saeys, 2010; Dietterich, 2000; Moult, 2005). A number of missense ensemble predictors, for example CONDEL (González-Pérez & López-Bigas, 2011), PONP (Olatubosun et al., 2012), Meta-SNP (Capriotti et al., 2013) and most recently REVEL (Ioannidis et al., 2016) have also been developed for the more usual task of binary classification, but as far as we are aware, this is the first use for quantitative prediction of missense impact.

I also performed several post-challenge analyses on the NAGLU dataset, examining the usefulness of structure information for identification of deleterious mutations and comparing the performance of the new ensemble method with other missense methods for binary classification. In the clinic, a major concern is not just to have an accurate predictor of pathogenicity, but also to assign a reliable probability that an assignment of pathogenic or benign is correct. The NAGLU challenge data set provided an opportunity for testing methods of assigning such probabilities on a clinically relevant dataset.

2.3 Methods

2.3.1 Challenge data and benchmark data

The challenge set of 165 *NAGLU* rare population missense mutations was provided by Jonathan H. LeBowitz (BioMarin). The SUMO-ligase CAGI challenge set was generated by the Fritz Roth lab using a competitive yeast complementation growth assay (Weile et al., 2017). Three sets of *UBE2I* (SUMO-ligase) mutations were provided – 1) a reliable (multiple measurements) set of 219 single missense mutations, 2) a less reliable set of 463 single missense mutations and 3) a set of 4427 double or more mutations per clone. The experimental NAGLU enzyme activity data and the SUMO-ligase yeast growth data were not released to CAGI participants until all predictions had been submitted. In addition, we also collected 90 *NAGLU* known disease-related variants from HGMD (Stenson et al., 2014), together with the 278 interspecies variants, as a benchmark set.

2.3.2 Data for training predictors of continuous activity

Methods training for both NAGLU enzyme activity and SUMO-ligase growth rates required data that are also on an appropriate continuous scale of biological activity (as opposed to the more usual pathogenic/benign classification). For this purpose, a set of enzyme activity data for 92 human Phenylalanine hydroxylase (PAH) variants from (http://www.biopku.org/pah/) was used, supplemented by a set of 139 PAH interspecies variants (identified by comparing the human sequence with those of seven PAH orthologs (HomoloGene, (NCBI Resource Coordinators, 2015)) with

sequence identities higher than 80%), assumed to have full activity. I also searched the literature for high throughput mutation datasets that might be appropriate for use as training data. Only one of these appeared suitable, a set of cell growth rate data for yeast ubiquitin (UBI4) mutations (Roscoe, Thayer, Zeldovich, Fushman, & Bolon, 2013). In practice, methods trained on these data performed poorly, and so its use was discontinued.

2.3.3 Combining multiple missense analysis methods to predict relative protein activity

For the ensemble methods, up to eleven missense analysis methods were used: Polyphen-2 (Adzhubei et al., 2010), SIFT (Ng & Henikoff, 2003), SNPs3D Profile (Yue & Moult, 2006), CADD (Kircher et al., 2014), Panther (Thomas et al., 2006), PON-P2 (Niroula et al., 2015), SNAP2 (Hecht et al., 2015), PROVEAN (Choi et al., 2012), VEST3 (Carter, Douville, Stenson, Cooper, & Karchin, 2013), LRT (Chun & Fay, 2009) and MutationTaster (Schwarz et al., 2010). The dbNSFP2.9 database (X. Liu, Jian, & Boerwinkle, 2013) was used to obtain CADD, PROVEAN, LRT, VEST3 and MutationTaster results. SNPs3D Profile results were obtained using the standalone in-house software. Results of other methods were obtained from the corresponding web-servers.

Binary predictions and associated scores were collected when both were available. Polyphen-2 ‘Probably damaging’ and ‘Possibly damaging’ were merged as a deleterious assignment. The MutationTaster deleterious set was compiled by combining the ‘A’ and ‘D’ categories, and the benign set consisted of the ‘P’ and ‘N’

categories. Four methods (CADD, SNPs3D profile, Panther, and VEST3) didn't directly report binary assignments. The recommended threshold score of 15 was used for CADD and the standard score threshold of zero was used for SNPs3D profile. A 'deleterious' score of 0.5 and a score of 0.77 were chosen as the cutoffs for Panther and VEST3 respectively, the values at which the distribution curves of deleterious and benign training sets crossed each other.

For machine learning based prediction of protein activity, two sets of input features were tested: One set consists of the score values returned by each of the 11 missense methods listed above. The other set consists of the binary assignments of benign or deleterious, represented as 0 or 1. Both feature sets also included the fraction of agreement (FOA) for a deleterious assignment across predictors, calculated as follows:

$$FOA = \sum_i C_i \Big/ \sum_i N_i$$

where the sum is over the number of missense methods included, and N_i is 1 if a binary assignment is available for the i-th method, and is 0 otherwise, C_i is 1 if the i-th method predicted deleterious and is 0 if the i-th method predicted benign or was not available.

Weka (Frank et al. 2016) with standard settings was used to test a number of machine learning models: logistic regression, linear regression, support vector machine (SVM) regression, multi-layer perceptron, M5 Rule, random tree and random forest. The overall best performance (as judged from Root mean square deviations (RMSD, see

supplementary methods), Pearson, and Spearman) on the PAH training set with 10-fold cross validation was returned for an SVM regression with an RBF kernel with the default settings and using the 11 method scores and FOA as features. However, the spread of performance across the best combinations of the feature sets and the ML methods was small (Pearson's r 0.84-0.87, RMSD 0.18-0.20, 10-fold cross validation) and so more extensive parameter optimization might have produced a different choice. In addition to the prediction of activity, CAGI4 rules also required estimated standard deviations for each activity value. I provided the RMSD on the PAH training set as the standard deviation for all predicted activities.

2.3.4 Scale calibration and manual adjustment for each challenge

The SVM regression model was used to predict the relative enzyme activity of each *NAGLU* mutation and the cell growth rate of each *UBE2I* (SUMO-ligase) mutation. Because the model was trained on a different gene (PAH) with enzyme activity measured using a different experimental assay, we expected some systematic bias in the predictions and assumed that results would require scaling for each challenge system. For NAGLU, a zero activity reference point was defined using 15 known disease mutations with reported zero enzyme activity (Beesley et al., 2004; Lee-Chen et al., 2002; Tessitore et al., 2000; Weber et al., 1999). A full activity reference point was defined by the 278 *NAGLU* interspecies variants compiled in the same way as the PAH interspecies variants described above. These reference points were used to linearly scale the NAGLU activity predictions. I also collected structural information on the NAGLU protein from SNPs3D stability (Yue et al., 2005) and FOLDX

(Guerois et al., 2002) predictions, as well as information on the functional role of individual residues from UniProt (UniProt Consortium, 2015). Two predictions affecting disulfide bonds were manually adjusted to 0.1 activity. Predictions for six residues were adjusted to lower predicted activity in an *ad hoc* manner, on the basis of predicted structure destabilization. The experimental data later showed that these manual adjustments did not improve overall prediction accuracy, and increased prediction error for three of the six residues. For SUMO-ligase, the distribution of experimental measurements was provided as part of the challenge. Two submissions were made using different calibration procedures. For the first, I used the closest experimental values to 0 and 1 as the zero and full growth rate reference points and applied a linear scaling procedure like that used for NAGLU. In the second submission, each predicted growth rate was uniquely matched to the corresponding ranked experimental value. We noted that the experimental distributions have a number of mutations with growth rates significantly higher than wild-type. For each challenge set, for the submission not mapped to the distribution of experimental data, it was necessary to reassign some growth rates to values greater than wild-type to match experiment. I increased the values for the top predicted growth rate subset, except for those that predicted destabilizing by SNPs3D Stability (Yue et al. 2005) and FOLDX (Guerois et al. 2002). I also took into account (Bernier-Villamor et al. 2002; UniProt Consortium 2015) several reports of mutations with enhanced growth rate. The experimental data showed that this procedure is less accurate than that without manual adjustments on most gain-of-function mutations (22 of 27 in set 1 and 47 of 52 in set 2). For Challenge set 3, where multiple mutations were present in each

sample, we assumed that the highest impact prediction dominated, and assigned that predicted value. The results of each challenge presented throughout the rest of the manuscript are based on a final set of predictions that include the manual adjustments.

All final predictions were adjusted to be 0 if below 0, as required by the CAGI4 submission instructions.

2.3.5 Positive and negative controls

Positive and negative control models were used to further evaluate the continuous predictions of relative protein activity. The positive control model estimated the performance expected if the computational method were perfect so that the only discrepancies arose from experimental error. For this purpose, simulated experimental errors were randomly drawn from a Gaussian distribution using the reported experimental mean and standard deviation based on the experimental error for each mutation. The performance was averaged from 1000 repeats of this process. The negative control adopted the algorithm proposed by the CAGI SUMO-ligase assessor as follows:

$$\text{Prediction Score} = \ln(\frac{P_m}{Q_m}) - \ln(\frac{P_w}{Q_w})$$

Where P_w and P_m are the probability of the wild type and mutated residue type occurring at the mutated position in a multiple sequence alignment and Q_w and Q_m are the background frequencies of the wild type and mutated residue respectively in the entire sequence profile.

2.3.6 Analysis of the influence of training set type and size on performance

The continuous value prediction models used a small training set of mutations and that set was from an unrelated protein. Once the submissions were made and the experimental data were available, for each of the challenges, I tested the influence of these factors as follows. 15% of the data was set aside for testing and a series of subsets of different sizes were randomly selected from the remainder. The machine learning model was retrained on each of these subsets. The procedure was repeated 10 times, omitting a different 15% data each time. Performance was then evaluated as a function of training set size.

2.3.7 Training and testing data for the binary predictor

For training ensemble binary predictors of pathogenicity, all mutations in an earlier version of HGMD (Stenson et al., 2003) were used as true positives and a set of interspecies variants were used as true negatives ('benign' mutations), compiled by comparing homolog protein sequences across species with at least 90% sequence identity over at least 80% of the full length (Yue & Moult, 2006). For testing pathogenicity models and assessing prediction reliability, I compiled two independent test data sets. The first set is composed of ClinVar (Landrum et al., 2016) variants with pathogenic or benign assignments, excluding all that are in HGMD (2014 version) (Stenson et al., 2014) and OMIM (http://omim.org/) in order to ensure independence from the commonly used training data. ClinVar 'likely pathogenic', 'likely benign' entries, and entries with conflicting ClinVar assignments were not included. The second is the challenge set of 165 *NAGLU* rare population missense

mutations. A complication in this analysis is choosing an activity level below which all mutations are pathogenic (that is, penetrance is 100%). In other data referenced by the data provider, pathogenic mutations are found at activities up to 45% but most are below 15%. Because of this uncertainty, I evaluated methods performance using both 10% and 30% relative enzyme activity cutoffs for pathogenicity.

2.3.8 Pathogenicity prediction models

Three machine learning methods were tested for binary state (pathogenic/benign) prediction models: Logistic Regression (Weka), Random Forest (Weka) and SVM (RBF kernel, SVMlight (Joachims 1999)). Features sets were the same as those used for continuous value prediction except that Panther and SNAP2 predictions were removed due to the difficulty of collecting the large number of predictions required from the corresponding web-servers. Models were trained using the HGMD dataset with default parameters. REVEL (Ioannidis et al., 2016) predictions were downloaded from (https://sites.google.com/site/revelgenomics/). The dbNSFP2.9 database (X. Liu et al., 2013) was used to map REVEL results to individual protein mutations.

2.3.9 Measuring prediction reliability

In the clinic, variants are often accepted as pathogenic or benign if the confidence in that assignment is estimated as greater than some threshold, typically 90%. For each binary prediction method, I therefore I evaluated the fraction of

variants that were predicted with reliability (PPV, positive predictive value, see supplementary methods) at 95%, 90%, 85% and so on. To this end, for each method, the data were sorted by the associated prediction score, from highest confidence score to lowest. For prediction of pathogenicity, the fraction of highest confidence variants with a given PPV was then determined. The resulting fractions versus PPV curves were plotted using R ggplot2 (Wickham H, 2009)(Wickham 2009). To reduce noise, the NAGLU dataset was expanded to 1000 variants by bootstrapping, and assessed by averaging over 1000 bootstrappings.

2.3.10 Measures of performance

For predicted NAGLU enzyme activity and SUMO-ligase yeast growth rate, I calculated the root mean square deviation (RMSD) as follows:

$$RMSD = \sqrt{\sum_{i} (X_{pred}^{i} - X_{exp}^{i})^2 / N}$$

Where X_{pred}^{i} and X_{exp}^{i} are the predicted and experimental value of the i-th mutation. I used in-house programs and EXCEL2013 to calculate the Pearson's r and the Spearman's rho. The true positive rate and false positive were defined as following:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

The area under the receiver operating characteristic (ROC) curves (AUC) were approximated using the R package pROC (Robin et al., 2011). AUCs of different methods were compared using DeLong's test (DeLong, DeLong, & Clarke-Pearson,

1988). When testing on the HGMD training set, I performed 10-fold cross validation. For evaluation of the accuracy of probability of pathogenicity estimates, the positive predicted value (PPV) is defined as follows:

$$PPV = \frac{TP}{TP + FP}$$

2.4 Results

2.4.1 Comparison of predicted and experimental enzyme activities

Figure 2-1A shows a scatter-plot for the NAGLU challenge mutations showing the relationship between all predicted and experimental enzyme activities. The overall RMSD between predicted and experimental values is 0.31, Pearson's r is 0.55, and Spearman's rho is 0.57. These values are worse than the cross validation results on the PAH training data, which are RMSD of 0.20, Pearson's r of 0.82 and Spearman's rho of 0.78. The NAGLU predicted values are also substantially worse than the positive control 'perfect prediction' RMSD of 0.12, 0.95 Pearson's r and 0.94 Spearman's rho (based on the reported experimental standard errors). There are a small number of serious outliers, and as the plot shows, most of these correspond to mutations identified by the assessor as 'hard to predict' on the basis of poor performance by all the top methods. A breakdown of performance by location in the structure (Figure 2-2) shows striking variations for the Pearson's correlation coefficient of 0.83, 0.50 and 0.39 for buried, partially exposed and surface mutations respectively. (Variant location based on the STRIDE (Eisenhaber & Argos, 1993; Eisenhaber, Lijnzaad, Argos, Sander, & Scharf, 1995; Frishman & Argos, 1995)

relative surface accessibility: buried core (≤ 0.05), partially exposed (> 0.05, ≤ 0.25) and surface (> 0.25)). The most serious outliers for both under and over-prediction of activity are in the partially or completely exposed subsets. Performance metrics are substantially improved omitting these ten, with RMSD of 0.24, Pearson's r of 0.71 and Spearman's rho of 0.71 (Table 2-1).

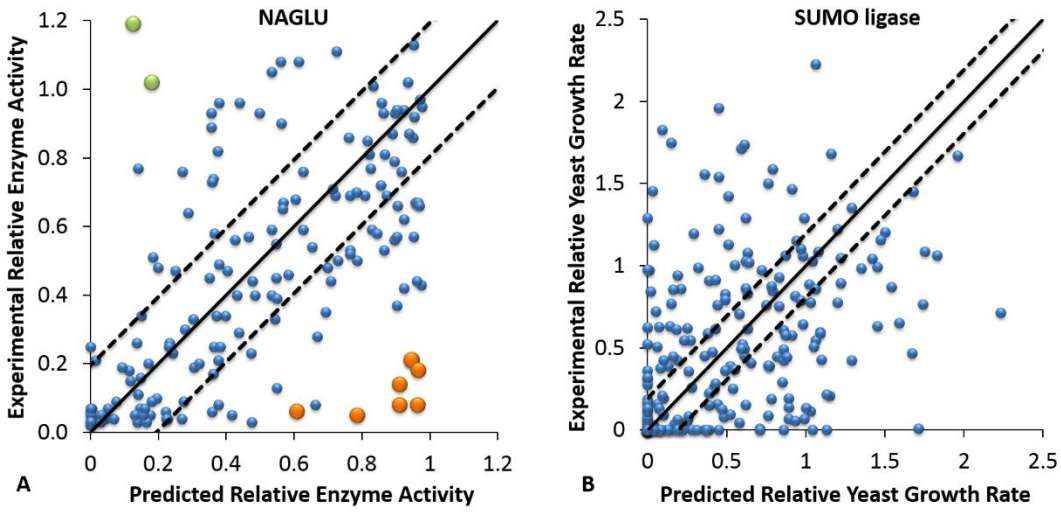


Figure 2-1. Prediction results for *NAGLU* and *UBE2I* (SUMO-ligase) mutations.

Figure 2-1A. Scatter-plot comparing experimental NAGLU relative enzyme activities (Y-axis) with the predicted values (X-axis) for the CAGI challenge variant set. Dashed lines delineate the expected prediction RMSD from based on training results. 61% of the predicted values are within the range of the estimated RMSD, but a few mutations have very large deviations from the experimental measurements. The over-estimates shown in orange and the under-estimates shown in green are the ten mutations selected by the assessor as 'hardest' to predict. See text and Figure 2-2 and 2-3 for a discussion of these.

Figure 2-1B. Scatter-plot comparing experimental relative yeast growth rates with the mapped predicted values for the SUMO-ligase CAGI challenge *UBE2I* mutation Set 1. Dashed lines delineate the expected prediction RMSD from the training on phenylalanine hydroxylase mutations. The correlation with experiment is substantially weaker than for the NAGLU challenge (Figure 2-1A). 39% of the predicted values are within the range of the estimated RMSD.

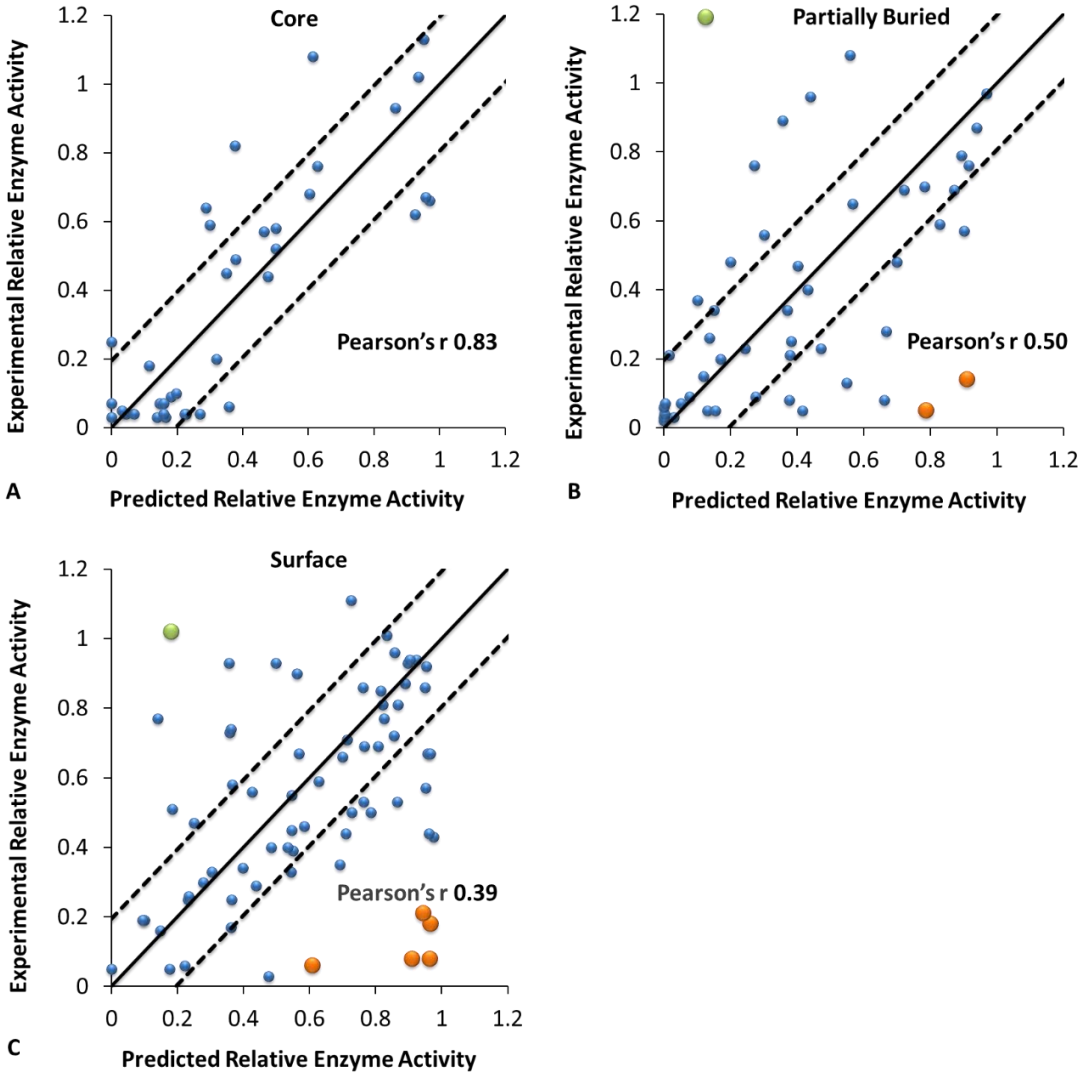


Figure 2-2. Scatterplot comparing experimental CAGI NAGLU relative enzyme activities with predicted values for three categories of surface accessibility. **2-2A**) core residues, **2-2B**) partially buried residues, **2-2C**) exposed residues on the surface. Dashed lines delineate the expected prediction RMSD expected from training performance. Predictions are most accurate in the core and least accurate on the surface. Orange and green colored points represent mutations considered 'hard' by the assessor. Performance is worst for surface residues and best for core residues,

Table 2-1. Metrics of prediction performance for NAGLU and SUMO-ligase

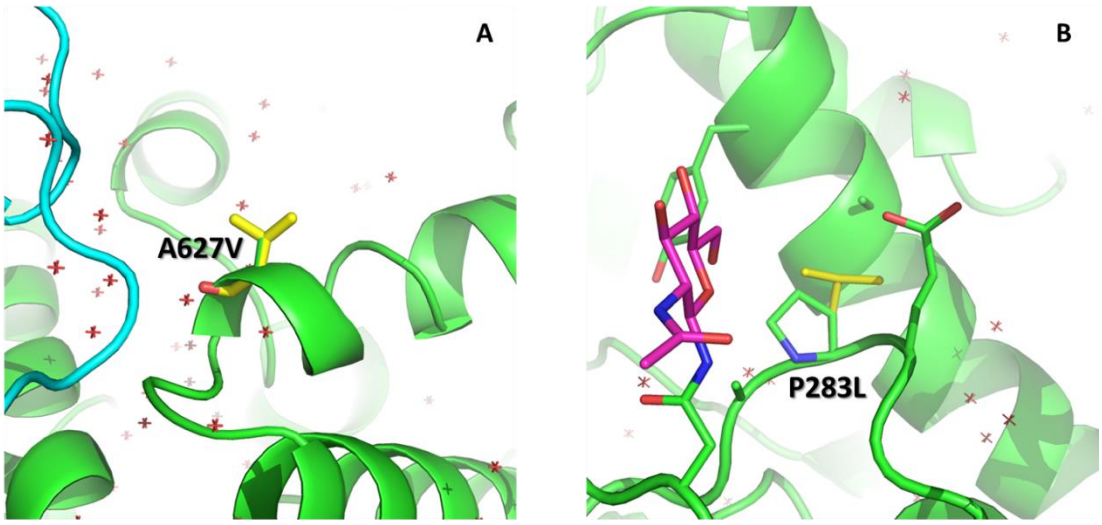
[a] In the SUMO-ligase challenge, we submitted two prediction sets, submission 1 using the scaled prediction scores (No Map), submission 2 (Mapped) mapping each predicted value to the experimental value of closest rank.

[b] In the positive control, I estimate the expected difference between experiment and prediction, given the reported experimental errors. That is, a perfect prediction method could not be more accurate than this. See MATERIALS AND METHODS.

[c] In the negative control, a prediction score was computed for each mutation based on amino acid frequency information only, using the equation described in MATERIALS AND METHODS. The resulting prediction scores were mapped to the experimental value of closest rank.

Challenge[a]		Prediction			Positive Control[b]			Negative Control[c]		
		RMSD	Pearson's r	Spearman's rho	RMSD	Pearson's r	Spearman's rho	RMSD	Pearson's r	Spearman's rho
NAGLU		0.31	0.55	0.57	0.12	0.95	0.94	0.42	0.45	0.48
NAGLU w/o outliers		0.24	0.71	0.71	0.14	0.92	0.93	0.39	0.53	0.57
SUMO - Ligase Set 1	No Map	0.55	0.39	0.46	0.24	0.91	0.92	0.59	0.30	0.38
	Mapped	0.55	0.39	0.46						
SUMO - Ligase Set 2	No Map	0.63	0.35	0.46	0.25	0.90	0.89	0.57	0.31	0.39
	Mapped	0.56	0.33	0.46						
SUMO - Ligase Set 3	No Map	0.57	0.21	0.20	0.26	0.89	0.82	0.57	0.24	0.22
	Mapped	0.59	0.18	0.20						

Are the ten outlier mutations cases where all the prediction methods systematically fail, or are these experimental artifacts of some sort? A definitive answer to this question is not possible without further experiments, but in some cases, likely explanations present themselves. For example, 10 out of 11 individual methods in the ensemble model and a structural method, SNPs3D Stability, predict mutation (*NAGLU* NP_000254.2:p.A627V) to be benign, but the reported experimental activity value is close to 0. Consistent with the prediction results, examination of a multiple sequence alignment shows A627 is at a variable position across species, where 15 different amino acid types are found. A627 is on the protein surface (Figure 2-3A) and the variant introduces a hydrophobic side chain (crystal structure from USPTO US08775146B2 (US08775146B2, 2014)). Under *in vivo* conditions, that may indeed have little impact, but in overexpression conditions of the experimental *in vitro* assay, aggregation may result. On the other hand, it is difficult to find any plausible explanation for some of the outliers. For example, one outlier (*NAGLU* NP_000254.2:p.P283L) is a partially exposed proline at an extremely conserved position (Figure 2-3B). All 11 individual prediction methods as well two structure-based methods, FOLDX (Guerois et al., 2002; Schymkowitz, Borg, et al., 2005) and SNPs3D Stability (Yue et al., 2005), predict this mutation deleterious. Inspection of the structure suggests no way in which the leucine side chain could be accommodated. The reported experimental activity is the highest of any of the variants, at 1.19.



High impact: 1 / 11 individual methods
SNPs3D Stability: +0.99 (Benign)
FOLDX ddg: +0.64 kcal/mol
Experimental relative activity: 0.08

High impact: 11 / 11 individual methods
SNPs3D Stability: -0.22 (Deleterious)
FOLDX ddg: +2.69 kcal/mol
Experimental relative activity: 1.19

Figure 2-3. Structural view of two of the 10 'hard' NAGLU outliers. Crystal structure is from USPTO US08775146B2. Mutated residues are yellow. Wild-type residues and environments in green, with a neighboring subunit in cyan. An N-Acetylglucosamine molecule is in magenta, ordered water molecules are red crosses.

2-3A 10 out 11 individual prediction methods and one structural method (SNPs3D Stability) predict A627V as benign, and it is a species variable surface residue, but reported experimental enzyme activity is only 0.08.

2-3B All 11 individual methods and two structural methods, FOLDX and SNPs3D Stability, predict P283L as deleterious and the proline is highly species conserved, but the experimental enzyme activity is 1.19.

Figure 2-1B is a scatter plot of the relationship between Submission 2 predicted and experimental growth rates for Set 1 *UBE2I* (SUMO-ligase) mutations. The performance is weaker (RMSD 0.55, Pearson's r 0.39, Spearman's rho 0.46) than the results for NAGLU, likely because of the complex relationship between aspects of SUMO-ligase function, its many substrates, and cell growth as well as effects from use of human protein in a yeast system. In contrast to NAGLU, the best performance is for surface residues (Pearson's r 0.59), and it is less good for mutations of buried (Pearson's r 0.35) and partially buried (0.29) residues. The results are worst for mutations in the substrate, SUMO, and SUMO-E3 ligase protein-protein interfaces ((Pearson's r 0.24, Figure 2-4). For example, in the experimental structure with a human SUMOylation substrate, RANGAP1 (PDB code 3UIP), the wild-type K74 forms a salt bridge with E526 of the substrate (Figure 2-5). Mutations (*UBE2I* NP_003336.1:p.K74S and *UBE2I* NP_003336.1:p.K74E) disrupt that interaction and in the case of K74E electrostatic repulsion is introduced. Both positions are conserved, and the mutations are overwhelmingly predicted deleterious, yet the experimental growth rates are higher than wild-type. On the other hand, mutation (*UBE2I* NP_003336.1:p.K74R) appears to enhance the salt bridge with E526, and four out of ten sequence methods and the two structure methods predict it as benign. Yet the experimental value shows complete loss of growth. At the CAGI meeting the data provider, Fritz Roth, agreed that a possible complication here is that interfaces between human SUMO-ligase and its human partners may have significantly different properties from the equivalent yeast interfaces, and that in general the substantial number of gain of function mutations may be due to this cause.

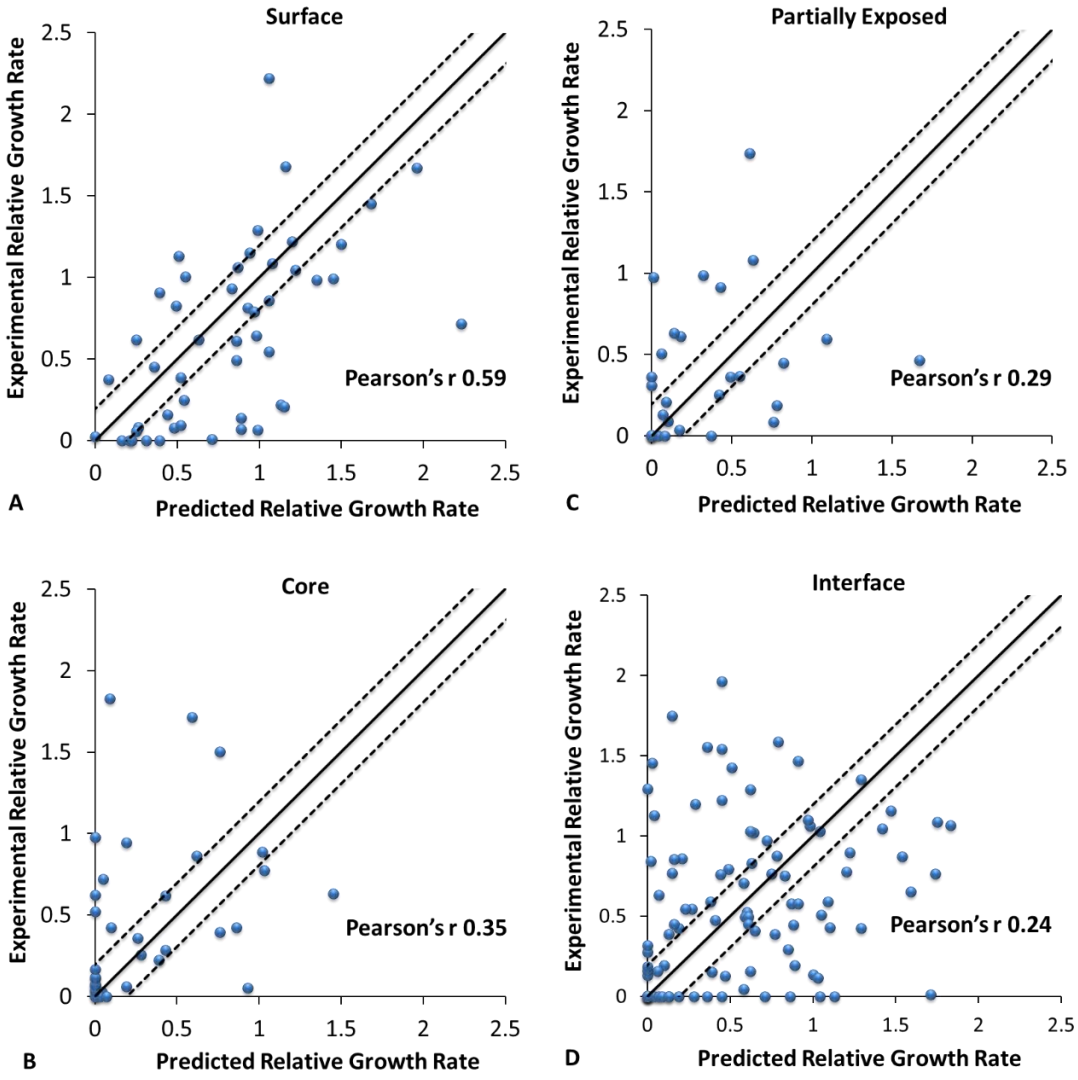


Figure 2-4. Scatterplots of the experimental SUMO-ligase set 1(Y-axis) relative cell growth rate versus predicted values. Mutations are divided into four categories based on solvent accessibility: **2-4A**) surface residues, **2-4B**) core residues, **2-4C**) partially exposed residues and **2-4D**) residues at the interfaces to SUMO, SUMO E1 ligase and SUMO E3 ligase. The dashed lines delineate the expected RMSD from the training on phenylalanine mutations. The best performance is for surface residues, and the worst is for protein interface residues.

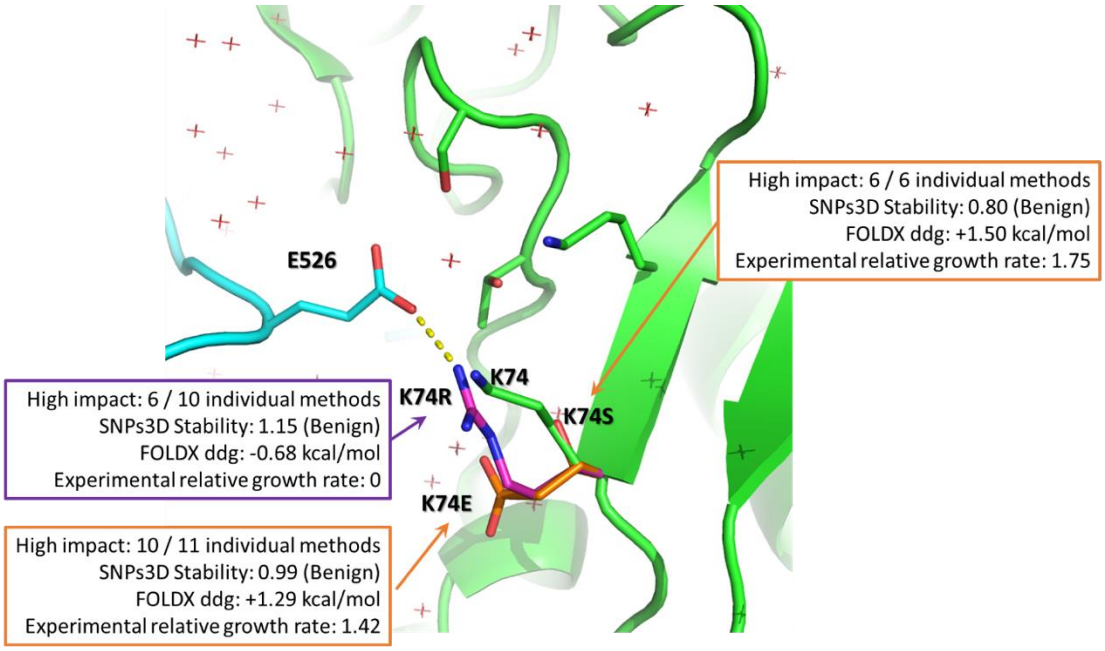


Figure 2-5. Structural view of three SUMO-ligase mutations of the same residue where predictions have large errors (PDB code 3UIP). Under-predicted mutation outliers (K74S and K74E) are orange and an over-predicted mutation outlier (K74R) is magenta. Wild-type residue K74 and its environment are green, and a SUMO-ligase substrate (RANGAP1) is cyan. K74R should make a stronger salt bridge to E526 than the wild-type K74 and consistent with that four individual sequence methods and two structure methods, SNPs3D Stability and FOLDX, predict this mutation as benign. The experimental value shows zero growth rate. On the other hand, most of the individual methods and FOLDX predict K74S and K74E deleterious and the structure shows these two mutations likely disrupt the contacts with the substrate residue. K74E may also induce electronic repulsion. But the experimental growth rates are higher than wild-type. Discrepancies for these and other interface-related mutations may reflect differences between human and yeast partner proteins structures or that different substrates have different modes of binding.

Table 2-2. Total number of variants in each dataset, and coverage of these by different prediction methods, for each dataset used. The SUMO-ligase set includes all non-redundant single mutations in CAGI challenge set 1, set 2 and set 3. ClinVar consists of 'pathogenic' and 'benign' missense variants excluding those also found in HGMD and/or OMIM.

At least N methods reporting	HGMD		ClinVar		NAGLU			SUMO-ligase		PAH	
N	Disease mutations	Inter-species variants	Pathogenic mutations	Benign variants	CAGI	Mutations in HGMD	Inter-species variants	CAGI	Inter-species variants	Disease mutations	Inter-Species variants
5	10006	10443	1582	2513	165	90	274	526	35	92	130
6	9956	10213	1582	2513	165	90	257	591	29	92	117
7	9296	9579	1582	2513	165	90	218	509	28	92	111
8	6354	7038	1561	2087	165	90	218	509	28	92	111
9	1460	2138	901	969	163	90	218	501	24	92	102
10					159	88	214	477	19	90	90
11					80	39	165	148	8	57	42
Total Count of variants	10865	13499	1584	2515	165	90	278	702	35	92	139

Some other SUMO-ligase substrates do not have exactly the same interface (Bernier-Villamor, Sampson, Matunis, & Lima, 2002). Thus, in general, it is not clear how altering the interface with one substrate may affect interactions with other substrates, and therefore what the overall effect on growth may be.

Table 2-1 summarizes all the agreement statistics between prediction and experiment for the *NAGLU* mutations and the *UBE2I* (SUMO-ligase) set 1, set 2 and set 3 mutations, together with the values for the positive and negative controls. (Data are for the SVM regression models described in Materials and Methods). Table 2-2 shows the number of missense analysis methods reporting for each data set. The results show our models outperformed the (quite sophisticated) negative control in the NAGLU challenge (RMSD 0.31 versus 0.42, Pearson's r 0.55 versus 0.45, and Spearman's rho 0.57 versus 0.48). The model is also effective on the SUMO-ligase set 1 (the most reliable single mutations) when compared to the negative control (RMSD 0.55 versus 0.59, Pearson's r 0.39 versus 0.30, and Spearman's rho 0.46 versus 0.38). The large gap between the method's performance and the positive control suggests that experimental error was likely not the limiting factor in the level of agreement with experiment.

2.4.2 NAGLU and SUMO-ligase challenge variant properties

The NAGLU challenge data are extracted from the ExAC database of population variants (Lek et al., 2016). In this respect it is a unique dataset – a set of variants found in a largely healthy population as opposed to the collections of known disease-

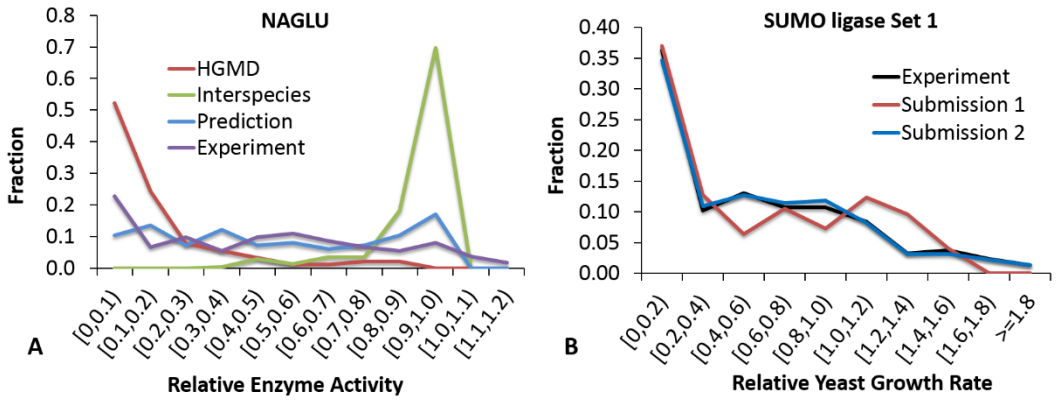
related mutations in databases such as HGMD (Stenson et al., 2003) and Clinvar (Landrum et al., 2016) and control sets of variants such as interspecies differences that are typically used for training and benchmarking methods. It is therefore of interest to ask how different the overall properties of these population variants are from the variants in the standard databases. Figure 2-6A shows that the predicted relative enzyme activity for the 90 *NAGLU* disease variants in HGMD and for 278 *NAGLU* interspecies variants have distinct distributions centered on 0 and 0.9~1 respectively, as expected. In contrast to this, the predictions for NAGLU CAGI challenge variants are approximately evenly distributed across the whole 0 to 1 range, in a manner similar to that of the experimental data.

Figure 2-6. Distributions of predicted and experimental enzyme activities

Figure 2-6A. Distribution of NAGLU relative enzyme activities 1) predicted for disease mutations in HGMD (HGMD, red); 2) predicted for inter-species variants (Interspecies, green); 3) predicted for mutations provided for the CAGI challenge (Prediction, blue), and 4) experimental activities for the challenge mutations (Experiment, purple). As expected, known disease mutations are predicted to have low activities and interspecies variant to have high activity. In contrast to these, the population variants have activities approximately equally distributed across the full range, for both prediction and experiment.

Figure 2-6B. Relative yeast growth rate distributions for *UBE2I* (SUMO-ligase) mutation Set 1. The distribution of the unmapped predicted values (Submission 1, red) only approximately matches the experimental distribution (Experiment, black),

available during the challenge. We submitted a second set of predictions in which each predicted value was mapped to the experimental value of closest rank (Submission 2, blue). This improves the overall match of the distributions (red and black) but not the prediction accuracy.



(Figure 2-6. See above for caption.)

Figure 2-6B shows a comparison of the distribution of predicted yeast growth rates for SUMO-ligase challenge Set 1 mutations compared to the experimental distribution. An unusual feature of the experimental distribution is a substantial number (19%) of gain of function mutations, and this resulted in a poor overall fit from our prediction model. For submission 1, the distribution at low growth rates (below 0.2) is close to experiment, but between 0.2 and 1.0 there are too few predicted values and there are too many moderate gain of function values (in the 1.0 to 1.4 range). The second submission, which mapped each predicted value to the closest experimental value, corrects these distribution errors and produced a better overall distribution but doesn't improve the prediction accuracy (Table 2-1). Set 2 showed similar results, whereas Set 3 shows many fewer gain-of-function mutations,

presumably because of the presence of multiple mutations in each sample (Figure 2-7).

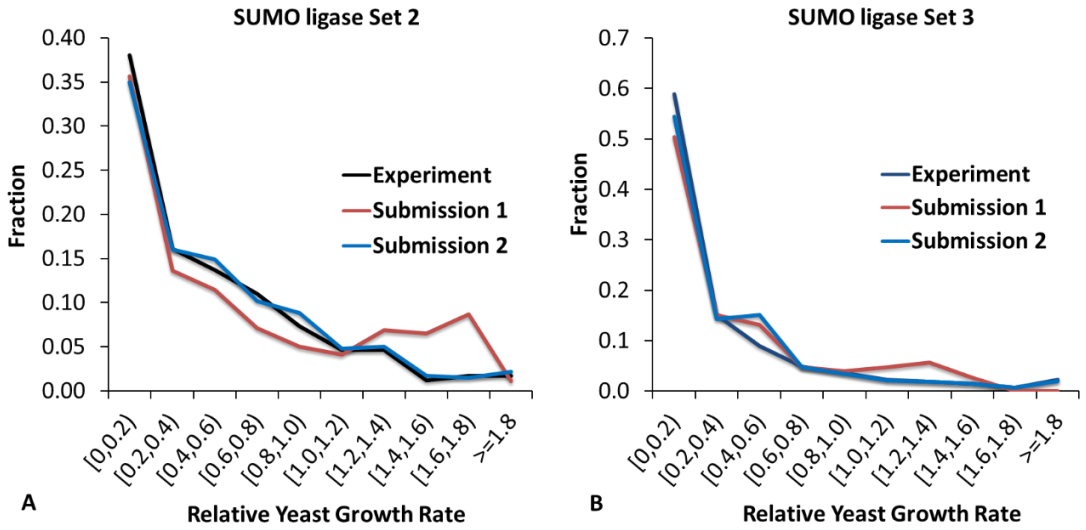


Figure 2-7. Distributions of experimental and predicted relative yeast growth rate distributions for the SUMO-ligase mutation Set 2 (**A**) and Set 3 (**B**). The distribution of the predicted values for our first submission (Submission 1, red) overestimates the number of experimental gain of function mutations (Experiment, black). The second submission (Submission 2, blue) corrects for this effect by mapping each value to the closest ranked experimental value. Set 3, with multiple mutations in each sample, has fewer gain-of-function mutants, as expected.

2.4.3 Role of structure destabilization

Thermodynamic destabilization of three-dimensional structure is established as playing a large role for monogenic disease-causing mutations (Yue et al., 2005), so it was of interest to examine what part this factor plays for the challenge variants. (This

analysis was undertaken after the results were known, and did not form part of our CAGI submissions). Figure 2-8A shows the distribution of destabilization scores from SNPs3D (Yue et al., 2005) for the NAGLU homo-trimer complex. At a NAGLU pathogenicity activity threshold of 0.3, a high fraction (68%) of the low activity variants are destabilizing, so, as in other monogenic diseases, this factor plays a major role.

The structure analysis is independent of the sequence methods and so provides some evidence for whether or not the 10 ‘hard’ predictions are experimental artifacts or systematic failures of the sequence methods. Two of the ‘hard’ variants with high experimental activity (*NAGLU* NP_000254.2:p.P283L and *NAGLU* NP_000254.2:p.G596C) are predicted destabilizing, consistent with the sequence analysis results and inconsistent with experiment. One of the ‘hard’ very low activity (0.06) variants ((*NAGLU* NP_000254.2:p.R377H), Figure 2-8B) is found to be destabilizing though, consistent with experiment and in disagreement with some sequence methods (5 out of 11). Wild-type R377 makes charge-dipole interactions with two main chain carbonyl groups (T343, A345) and a side chain hydroxyl group (Y335) so stabilizing a turn, and these interactions are absent for the variant (Figure 2-8B). The other seven ‘hard’ variants are all low activity and predicted to be not-destabilizing (lower right quadrant in Figure 2-8A). This could be because some other mechanism (for example involvement in catalysis) causes the low activity or because of experimental artifacts. Inspection of the structural environment does not reveal any such mechanisms, reinforcing the impression that these are experimental artifacts.

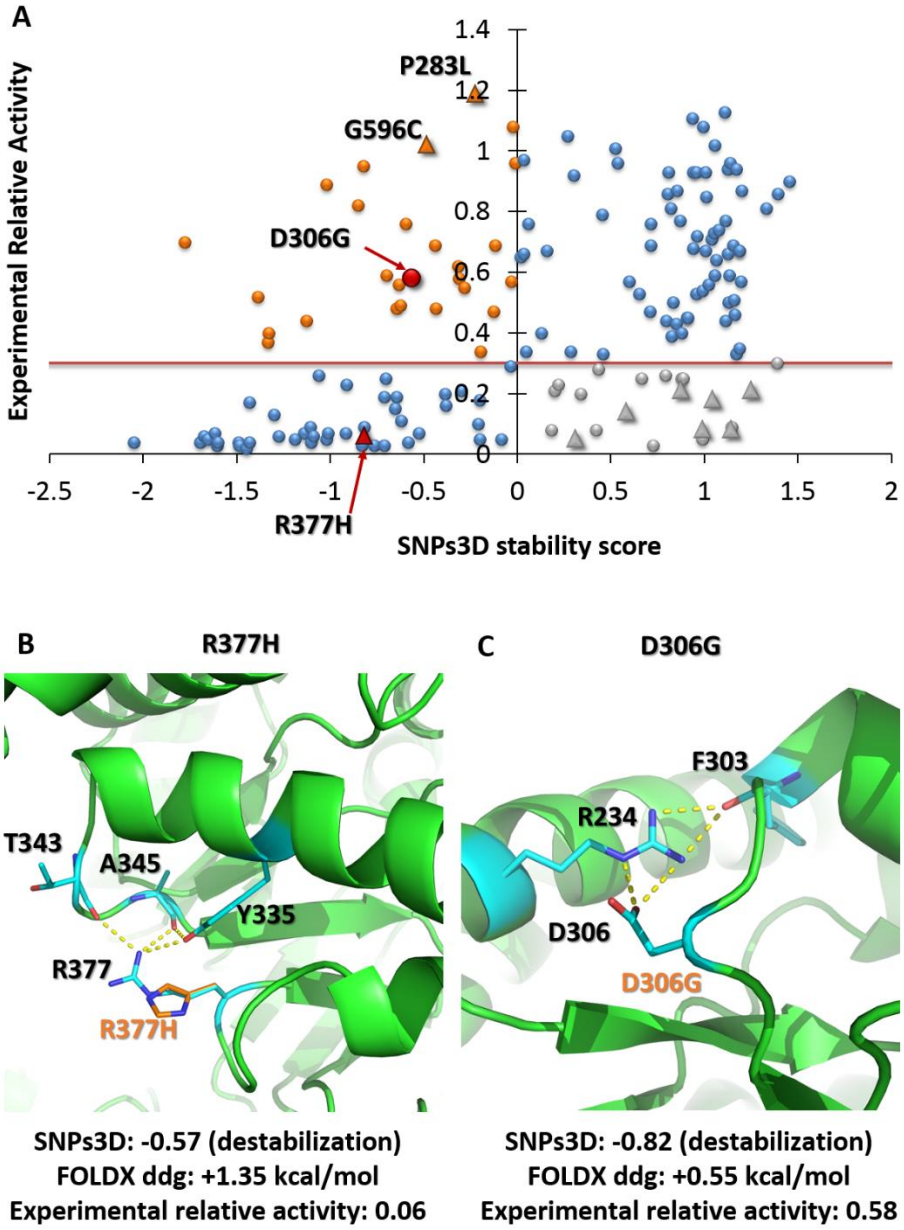
17% of the stability predictions disagree with the experimental data – predicted destabilizing but with higher than pathogenic activity. These partly reflect the shortcomings of present stability analysis methods as illustrated by the example of mutation (*NAGLU* NP_000254.2:p.D306G) (Figure 2-8C). Wild-type D306 forms electrostatic interactions with R234 that is absent for the variant. In reality, loss of this interaction is likely largely compensated for by increased solvation energy, a factor poorly represented in the SNPs3D model. There is scope for improvement of these methods in this and a number of other ways.

Figure 2-8. The role of thermodynamic destabilization in loss of function mutations.

2-8A) Scatter plot comparing SNPs3D stability scores with experimental relative enzyme activity of NAGLU. Blue point variants in the lower left quadrant (68% of all those with low (< 0.3) activity) are predicted to destabilize the structure. Those at the upper right are predicted not destabilizing, consistent with their high activity. Those at the lower right (gray) are predicted to have low activity for reasons other than destabilization. The upper left quadrant variants (orange) are predicted destabilizing even though the experimental activity is high. Triangles show the location of the ten 'hard to predict' variants.

2-8B) Structural context of NAGLU 'hard' outlier R377H (red in [A]). Predicted destabilization is consistent with the low experimental activity. A substantial fraction (5 out of 11) of sequence methods predict this variant to be benign.

2-8C) Structural view of variant D306G (red in [A]), predicted to be destabilizing, inconsistent with the experimental activity. Although the variant disrupts some electrostatic interactions, these are likely compensated by greater solvation. (Cyan: wild-type residues and interaction partners, orange: variants).



(Figure 2-8. See above for caption.)

2.4.4 Effect of training set size and choice of training data

One obvious drawback to my approach is the limited number (activities for 231 phenylalanine hydroxylase mutations) of training data. Further, training on that single system may introduce systematic bias. In order to evaluate whether the performance of the model is restricted by these two factors, I retrained using the NAGLU enzyme activity data, after these were released to the CAGI community (see Materials and Methods). A range of training set sizes was used to determine the contribution of that factor to accuracy. For each size, I retrained and measured performance, and averaged over 10 repeats. For each training, 15% of the data were randomly chosen for evaluation, and omitted from training. Figure 2-9 shows that performance converged rapidly as the size of the training set increased beyond 100 mutations, showing that training set used in the CAGI challenges was large enough and not a factor limiting accuracy. Comparison between the converged performance and the performance in the blind CAGI challenges showed only a slight improvement of 0.05 RMSD and 0.07 Spearman's rho for NAGLU and 0.08 RMSD for SUMO-ligase, so that the loss of performance from training on the phenylalanine hydroxylase system is small. Similar results were obtained for the SUMO-ligase challenge. Together, this analysis shows that the results were not substantially limited by either the training set size or training on a different system, and other factors must account for the worse than positive control performance.

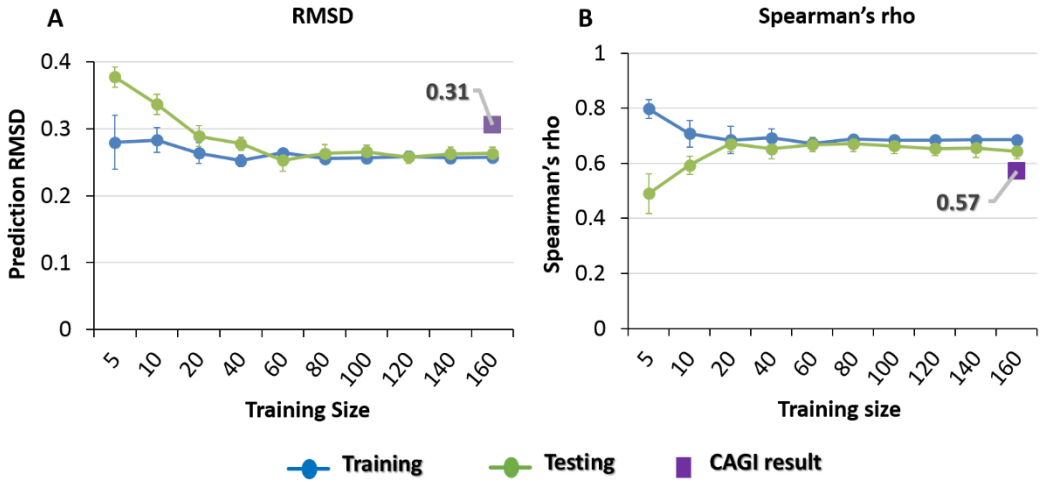


Figure 2-9. Blue: average training set performance, green: average test set performance (15% of data omitted from training). Averages over 10 runs. Purple rectangles show performance in the CAGI challenge with the model trained on PAH.

2-9A) RMSD, 2-9B) Spearman rank correlation coefficient. Prediction performance converges rapidly as the training set size increases beyond 100 mutations. Training on the target protein rather than Phenylalanine hydroxylase (PAH) only slightly improves performance (0.05 RMSD and 0.07 Spearman's rho). Thus, training set size and training on PAH are not limiting factors in performance.

2.4.5 Predicting pathogenicity using ensemble methods

Post-challenge, I also investigated how well ensemble methods perform on assigning pathogenicity in the clinically relevant NAGLU data, compared with performance on standard benchmarking datasets. For these binary predictions (pathogenic/not pathogenic), I trained ensemble methods based on nine individual predictors (CADD

(Kircher et al., 2014), LRT (Chun & Fay, 2009), MutationTaster (Schwarz et al., 2010), PON-P2 (Niroula et al., 2015), PPH2 (Adzhubei et al., 2010), PROVEAN (Choi et al., 2012), SIFT (Ng & Henikoff, 2003), SNPs3D Profile (Yue & Moult, 2006) and VEST3 (Carter et al., 2013)) with three machine learning models (Logistic Regression, Random Forest, and SVM). Training was performed on a version of HGMD (Stenson et al., 2003) and a set of interspecies variants (see Materials and Methods). Results were evaluated using 10-fold cross-validation. When tested on HGMD, the ROC curves and AUCs of the ensemble machine learning predictors show better performance than any of the individual methods, with the highest AUC of 0.98 (Figure 2-10A and Table 2-3), although most perform extremely well. A number of individual predictors are partially or completely trained on HGMD, so to control for this factor, I also tested on a subset of ClinVar variants not in HGMD or OMIM (another common source of training data). (Figure 2-10B and Table 2-3). Though still better than most individual predictors, my ensemble predictors (best AUC 0.95) were slightly but significantly outperformed by VEST3 (Carter et al., 2013) (AUC 0.96) and the new ensemble method REVEL (Ioannidis et al., 2016) (AUC 0.97). As Figures 2-10C and Table 2-3 show, when the same methods were tested on the more relevant challenge *NAGLU* variant set, all showed substantially deteriorated performance (AUC up to 0.84 for the ensemble methods, slightly better than any other tested methods). Relative performance is insensitive to the exact activity threshold for pathogenic loss of activity (Table 2-3). I also converted the continuous NAGLU activity predictions to binary assignments and generated a ROC curve. That results in an AUC of 0.82, with both 0.1 and 0.3 activity cutoffs. Evidently, the

distribution of activities found in the general population (all activities approximately equally likely to be encountered) are much more challenging for all methods than distinguishing between only pathogenic and interspecies variants.

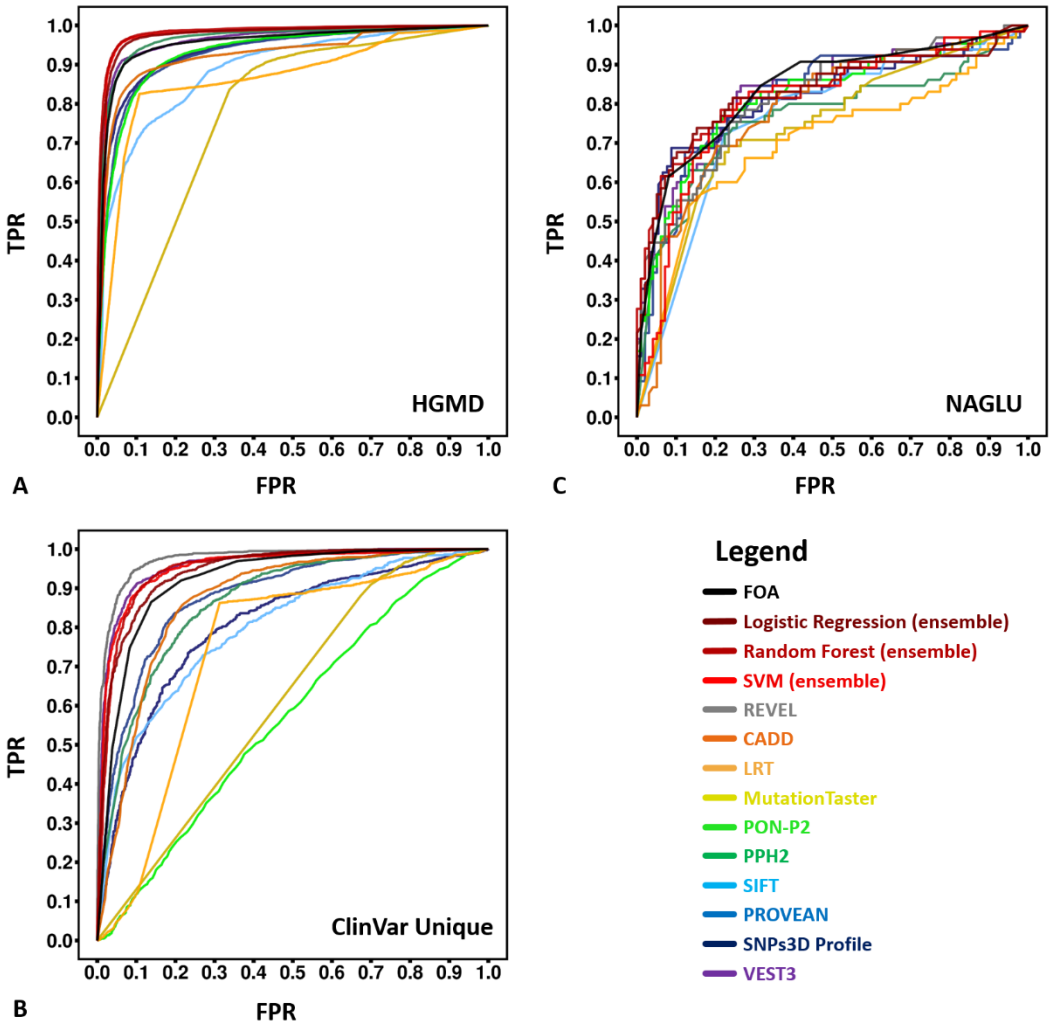
Figure 2-10. ROC (receiver operating characteristic) curves for predictions of pathogenicity by the new ensemble methods and other methods on HGMD, ClinVar unique and NAGLU challenge sets. For NAGLU, the pathogenicity threshold is an activity of 0.3 of wild-type. The AUC (area under curve) of these ROC curves are listed in Table 2-3.

2-10A For HGMD test data, the new ensemble models (Logistic Regression 0.98, Random Forest 0.98 and SVM 0.97) outperformed all constituent individual predictors on the HGMD test dataset. PPH2 and VEST3, which were also trained partially or completely on HGMD, have slight but significantly (P-value < 2.2e-6) worse AUCs.

2-10B For the unique ClinVar dataset (no overlap with HGMD or OMIM), another ensemble method, REVEL, outperformed all other methods. The next highest AUCs, for VEST3 and my ensemble models, are slightly but significantly (P-value < 0.05) smaller.

2-10C For the *NAGLU* rare population variants, all methods perform substantially worse than on HGMD and ClinVar. My ensemble FOA (fraction of agreement) method has the best AUC of 0.84, followed by my Logistic Regression and Random

Forest models, and VEST3. All four are not significantly different from each other (P-values > 0.05).



(Figure 2-10. See above for caption.)

Table 2-3. Metrics of binary prediction performance

Methods	HGMD			ClinVar			NAGLU[d]			NAGLU[c]
	AUC	Fraction 1[e]	Fraction 2[f]	AUC	Fraction 1[e]	Fraction 2[f]	AUC	Fraction 1[e]	Fraction 2[e]	AUC
Logistic Regression[a]	0.98	0.89	0.75	0.94	0.72	0.90	0.83	0.38	0.06	0.83
Random Forest[a]	0.98	0.90	0.75	0.95	0.78	0.93	0.83	0.43	0	0.83
SVM[a]	0.97	0.90	0.75	0.95	0.82	0.93	0.81	0.09	0.18	0.79
FOA[b]	0.96	0.85	0.73	0.92	0.27	0.88	0.84	0.28	0.56	0.83
REVEL	0.96	0.77	0.61	0.97	0.90	0.96	0.82	0.37	0.22	0.82
CADD	0.93	0.78	0.64	0.87	0.01	0.79	0.79	0.01	0	0.82
LRT	0.86	0.61	0.18	0.74	0	0	0.70	0	0.01	0.68
MutationTaster	0.77	0	0	0.61	0	0.22	0.74	0	0	0.79
PON-P2	0.93	0.65	0.54	0.57	0	0.03	0.82	0.22	0	0.82
PPH2	0.97	0.84	0.69	0.86	0.07	0.58	0.76	0	0	0.77
PROVEAN	0.93	0.76	0.63	0.88	0.26	0.68	0.82	0.08	0.50	0.82
SIFT	0.89	0.20	0.37	0.80	0	0.27	0.76	0	0	0.79
SNPs3D Profile	0.94	0.73	0.46	0.80	0.03	0	0.82	0.33	0.01	0.80
VEST3	0.96	0.85	0.74	0.96	0.84	0.94	0.83	0.30	0.44	0.83

[a] Ensemble model combining nine individual missense mutation analysis methods

[b] Fraction of the nine methods making a deleterious assignment

[c] Using NAGLU relative activity cutoff of 0.1

[d] Using NAGLU cutoff of 0.3

2.4.6 Reliability of pathogenic assignments

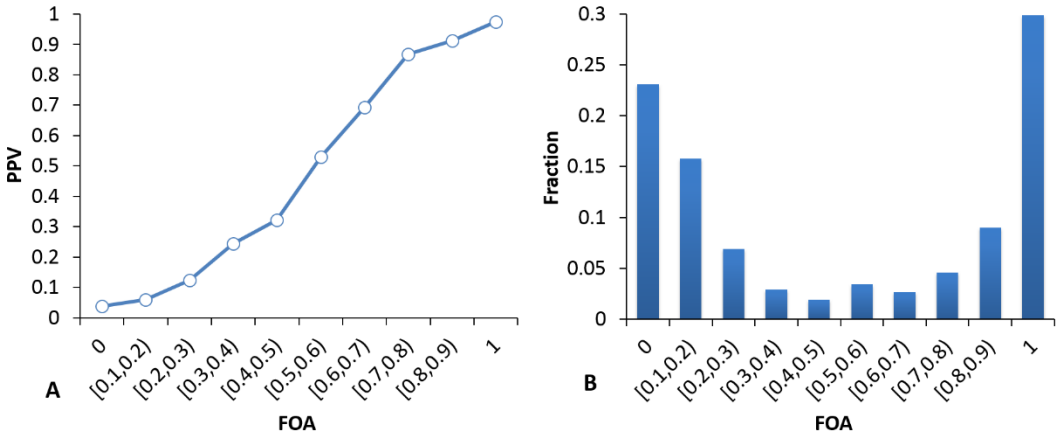
I investigated the effectiveness of ensemble methods for estimating the reliability of pathogenic assignments using the results from the binary pathogenicity analysis described above. To examine whether there is a useful ensemble signal to be exploited, I first examined the PPV as a function of the fraction of methods agreeing on a deleterious assignment (FOA) for the HGMD and interspecies dataset. Table 2-2 shows the number of methods included. There is strong dependence of PPV on FOA with the HGMD set (Figure 2-11A): For the set of variants where all nine methods predict deleterious, the PPV is 0.97 and the PPV is above 0.9 even when only 7 out of 9 methods predict deleterious. At the other end of the scale, the PPV is 0.04 when no method predicts deleterious and still below 0.1 even where two methods predict deleterious, so that in all 78% of mutations have better than 90% confidence assignments of either pathogenic or benign (Figure 2-11B). Thus even a very simple ensemble method shows promise for this purpose.

Figure 2-11 Initial results of estimating assignment reliability

2-11A. Relationship between the fraction of methods that agree on a deleterious assignment (FOA) and the positive predictive value, PPV (fraction of predicted pathogenic variants that are pathogenic), for HGMD and interspecies variants.

2-11B. Fraction of variants in each bin. Approximately 39% of variants can be predicted pathogenic with 90% or greater confidence (PPV) and 39% can be predicted benign with 90% or greater confidence (NPV). This simple analysis

demonstrated a potential usefulness of ensemble methods in assigning prediction reliability.

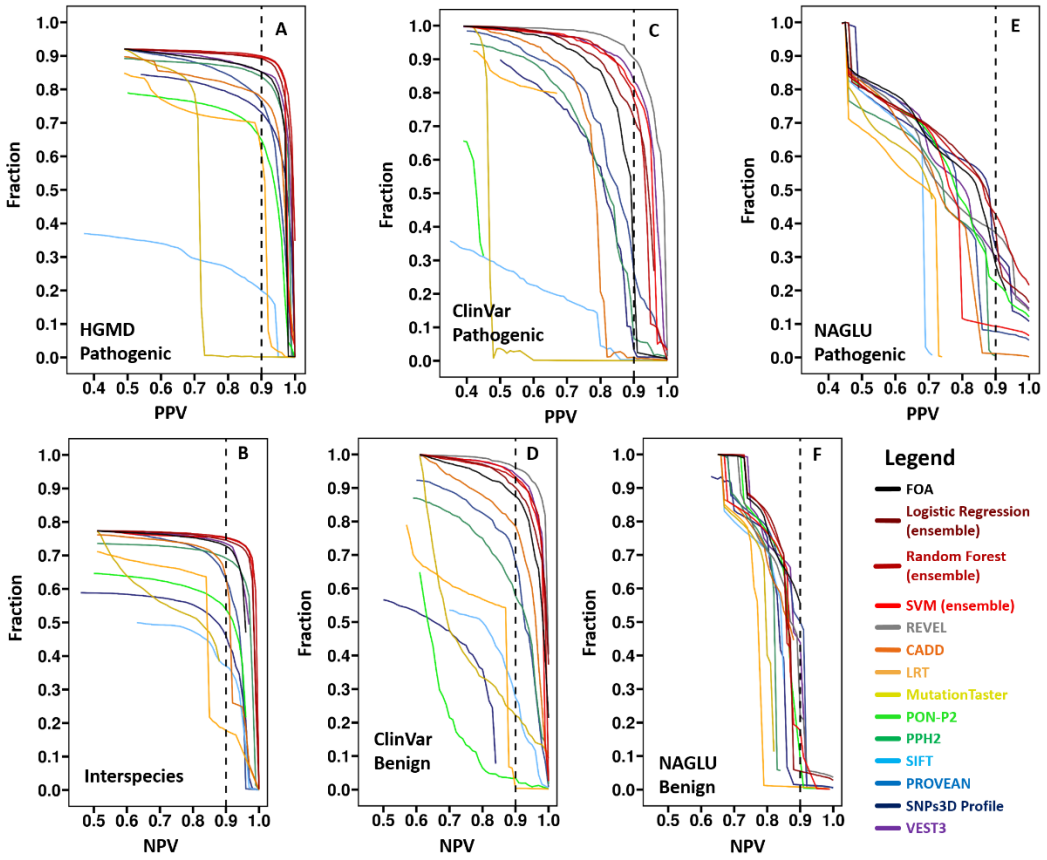


(Figure 2-11. See above for caption.)

A fuller analysis is shown in Figure 2-12. Here the fraction of variants meeting a given reliability threshold is plotted as a function of the threshold, for both confidence in pathogenicity (top panels) and non-pathogenicity (bottom panels). As with the pathogenicity assignment results above, my ensemble methods and REVEL perform best on the HGMD and ClinVar sets respectively. Also as with the pathogenicity assignment, performance is substantially better on the HGMD and ClinVar test sets than on the NAGLU data. For HGMD, the best methods assign pathogenicity with 90% or greater confidence for 90% of the data, and benign assignments with equal confidence are made for about 75% of data. Pathogenicity confidence on the ClinVar set is similar, with a higher fraction meeting 90% confidence criterion (96%) for benign assignments. For the more realistic NAGLU dataset using an activity of 0.3 as the pathogenicity threshold, 43% of the pathogenic variants are predicted with 90% or

better accuracy, and 56% benign assignments are 90% or better correct. However, the dependence of accuracy on the threshold is steep for both these numbers, and precise values are likely to be dataset specific. Overall, the results do show that ensemble methods are advantageous for assigning reliability to pathogenicity assignments, and that the fraction of variants for which 90% confidence can be reached in the clinic is likely quite high. More realistic datasets such as the NAGLU one are needed to further investigate these properties.

Figure 2-12. Fraction of data for which pathogenicity or benign status is predicted at a specified level of confidence, as a function of the confidence level, for HGMD (2-12A, 2-12B), ClinVar (2-12C, 2-12D) and the NAGLU challenge dataset (pathogenicity cutoff of 0.3, 2-12E, 2-12F). Vertical dashed lines show the 0.9 reliability threshold. For each dataset, the top panel shows the fraction of pathogenic variants meeting a reliability (PPV) threshold as a function of threshold and the bottom panel shows the equivalent data for reliability of benign assignment (NPV). My ensemble methods and REVEL perform best on the HGMD and ClinVar sets respectively. Overall, even in the demanding NAGLU dataset, a substantial fraction of variants can be assigned as pathogenic or benign with high confidence.



(Figure 2-12. See above for caption.)

2.5 Discussion

2.5.1 Ensemble methods for the NAGLU and SUMO-ligase challenges

The NAGLU and SUMO-ligase challenges are unusual in that CAGI participants were asked to predict a continuous variable – in the case of NAGLU, relative enzyme activity, and in the case of SUMO-ligase, relative growth rate in a yeast complementation assay. Most missense analysis methods are designed to make a binary assignment of pathogenic or non-pathogenic, and so are not immediately applicable to the challenges. To address this, we explored the use of an ensemble

strategy, incorporating up to 11 of the binary assignment methods. Ensemble methods have already been shown to be effective for the binary pathogenicity assignment task (Capriotti et al., 2013; González-Pérez & López-Bigas, 2011; Ioannidis et al., 2016; Olatubosun et al., 2012). Here we assume that the more single methods make a pathogenic assignment for a given variant, the lower the corresponding protein activity will be. As the simple FOA (fraction of agreement between methods) approach demonstrates, this is the case. Use of confidence scores for each contributing method rather than binary values makes the procedure more nuanced, and machine learning provides a means of combining the methods in a balanced way. A potential limitation was the lack of suitable enzyme activity training data, but post-challenge analysis showed that as few as 100 phenylalanine hydroxylase variant activities were sufficient, and also that there was no significant bias from training on that system. The ensemble approach was successful in that it performed well, although it was slightly behind the best performers. In the NAGLU challenge, the ensemble approach was marginally outperformed by MutPred2 (unpublished, -0.005 in RMSD, +0.05 in Pearson's r, +0.04 in Spearman's rho and +0.00 in AUC) and by Evolution Action (Katsonis and Lichtarge 2014, -0.028 in RMSD, +0.001 in Pearson's r, -0.019 in Spearman's rho and +0.03 in AUC). In the SUMO ligase challenge, our two submissions of the ensemble approach performed best on set 1 and set 2 respectively, but were outperformed by most other methods on set 3 (multiple mutation set), probably due to the assumption that growth would be determined by the most deleterious mutation for each sample, rather than affected additively. For data set 1, I also compared prediction performance for mutations at positions where

residue types are identical in human and yeast with that at positions where the residues are different. The former subset contains a relatively larger proportion of mutations with low relative growth rates. For example, there are more than twice as many zero growth mutations where the wild-type human and yeast residues are identical as opposed to different. But the performance of the ensemble method was not sensitive to this data partition (difference in RMSD ~0.07, in Pearson's r ~0.06, and in Spearman's rho <0.005). There is little difference between the fractions of wild-type identical residues versus non-identical for the subset of mutations with relative growth rates greater than 1.0. Neither our ensemble approach nor other best performers provided revolutionary accuracy. As discussed below, limitations in all contemporary approaches probably ensure that is not possible.

2.5.2 Accuracy

Although the methods used here and others in CAGI produce very strong statistical significance in terms of the relationship between predicted and experimental activity values, the agreement appears substantially less than expected, given the reported experimental accuracy. What limits the accuracy? – Some part of the disagreement may be due to experimental artifacts. For example as noted earlier, for one of the 10 *NAGLU* 'hard' variants the conditions of expression in the cell line may contribute to aggregation not encountered *in vivo*. For SUMO-ligase, as discussed in Results, differences between yeast and human proteins contribute to discrepancies.

Overall though, most of the discrepancy likely comes from the inherent deficiencies of the methods. Nearly all primarily attempt to relate sequence conservation patterns to pathogenicity (some also incorporate partial structure information (Adzhubei et al., 2010; Carter et al., 2013; Hecht et al., 2015)). Although there clearly is a qualitative relationship of this type, there is no theoretical framework providing a quantitative relationship. Such a framework would need to relate phylogenic profiles to fitness, something which the molecular evolution community has not succeeded in doing after many years of effort (Orr, 2009). Further, the relationship between fitness and disease relevance is also not straightforward. As a consequence, all current pathogenicity prediction methods are *ad h*oc, using calibration or machine learning to achieve some level of quantitation. Given that, they are surprisingly effective. There are a number of ways in which accuracy may improve in the future. In my results, there is markedly different accuracy for surface and interior residues, so that treating these classes of residues differently may be useful. Other structural and functional information may also help. Specific training only on variants where individual methods do not correlate well might be helpful, if there are sufficient data and an appropriate algorithm for training. More generally, at present, most methods are completely non-specific, and are applied to different proteins without incorporating information pertinent to each case. In future, we envision that protein specific models will be built. There is also a major requirement for more realistic training and testing datasets, such as NAGLU.

2.5.3 Assigning pathogenicity

As noted earlier, the NAGLU challenge data set is so far unique in that it consists of protein activity data drawn from a background population representative of that expected in the clinic. The commonly used HGMD and ClinVar databases, although useful compilations of clinically relevant data, are usually paired with highly benign controls for training and testing purposes, and so not very representative of clinical encounters. Therefore, I also tested an ensemble approach for assigning pathogenicity in the NAGLU dataset, compared to standard benchmarks. The new ensemble method and many others tested here perform extremely well on two standard benchmark sets, HGMD (Stenson et al., 2014) and a unique subset of ClinVar (Landrum et al., 2016), many with AUCs of over 95%. Both my ensemble method and another recent ensemble approach, REVEL (Ioannidis et al., 2016) have relatively good performance on the NAGLU data, but overall, all methods are strikingly less effective (best AUCs up to 0.84). The results suggest that we need many more clinically relevant datasets like NAGLU in order to realistically evaluate the pathogenicity assignment methods.

2.5.4 Utilization of protein structure information

As demonstrated here and in other work (Adzhubei et al., 2010; Baugh et al., 2016; Carter et al., 2013; Folkman et al., 2016; Hecht et al., 2015; Redler et al., 2016; Yue et al., 2005), analysis based on protein structure provides an orthogonal approach that, in spite of its own accuracy limitations, can sometimes provide valuable insight into the atomic level mechanisms in play. In particular, as with other monogenic disease-related mutations (Yue et al., 2005), for NAGLU, structure analysis shows a large

fraction operate by destabilizing protein three-dimensional structure. There is considerable scope for further improvement of these approaches, using more biophysical approaches (Seeliger & de Groot, 2010).

2.5.5 Reliability for pathogenicity assignments

In the clinic a major concern is not just to have an accurate predictor of pathogenicity, but also to be able to have a reliable probability that an assignment of pathogenic or benign is correct: a method may be highly accurate some of the time and fail on a subset of variants, and it is important to know when the prediction can be trusted and with what confidence. Because of a lack of well-tested reliability estimates, present clinical guidelines allow computational methods of predicting pathogenicity only secondary status as evidence for establishing a genetic cause for disease symptoms (Richards et al., 2015). The challenge NAGLU data set provided an opportunity for testing methods of assigning such probabilities on a clinically relevant dataset. The ensemble methods reported here, as well as other ensemble approaches such as REVEL (Ioannidis et al., 2016), are among the best for this purpose. Encouragingly, even on the realistic *NAGLU* population variants, a substantial fraction (up to 40%) of pathogenicity assignments can be made with greater than 90% confidence. More testing on diverse mutation sets is needed to establish clinical applicability.

2.6 Acknowledgements

This work was supported in part by NIH R01GM104436 and R01GM120364 to JM. The CAGI experiment coordination is supported by NIH U41 HG007446 and the CAGI conference by NIH R13 HG006650. We are grateful to the NAGLU (Jonathan H. LeBowitz, Wyatt T. Clark and G. Karen Yu) and SUMO ligase (Fritz Roth) dataset providers for making these challenges possible.

Chapter 3: Characterizing and comparing missense variants in monogenic disease and in cancer

3.1 Introduction

3.1.1 Overview

The large amount of genomic data now available for monogenic disease and for cancer has vastly expanded our knowledge of which mutations are involved in these diseases (Martincorena & Campbell, 2015; Shendure & Akey, 2015). In monogenic disease, over 7000 monogenic diseases and over 10,000 related genes have been described in the Online Mendelian Inheritance in Man (OMIM) database (http://omim.org/). In HGMD (Stenson et al., 2014), there are over 2,800 genes where some monogenic disease-causative mutations have been identified, over 50% of which are missense mutations. Sequencing of over 20,000 cancer sample exomes and a growing number of complete cancer genomes has revealed the mutation landscape for dozens of cancer types (Martincorena & Campbell, 2015; Vogelstein et al., 2013). Most of these data are available through three large consortia, the Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/), the Catalogue of Somatic Mutations in Cancer (COSMIC, http://cancer.sanger.ac.uk/cosmic) and the International Cancer Genome Consortium (ICGC) (http://icgc.org). The mutation load found varies by more than two orders of magnitude among individual samples as well as by cancer type (Martincorena & Campbell, 2015; Vogelstein et al., 2013). For example, acute myeloid leukemia and some pediatric cancers may carry less than 10 nonsynonymous somatic mutations per tumor, while exogenous mutagen induced cancers such as lung

cancer and melanoma typically have an average of around two hundred (Alexandrov et al., 2013; Lawrence et al., 2014; Vogelstein et al., 2013). It has been generally accepted that only a small number of the somatic mutations (4-6 (Armitage & Doll, 1954; Sabarinathan et al., 2017)) (the 'drivers') in each sample are responsible for the development of the disease. A recent comprehensive study estimates the average total number of driver mutations per sample as 4.6, including both SNVs and CNVs (Sabarinathan et al., 2017).

A variety of mutation types may be causative of monogenic diseases or be cancer drivers, including single base changes resulting in effects on expression and splicing, amino acid substitutions (missense) and premature stop codons, as well as small insertions and deletions (Ciriello et al., 2013; Stenson et al., 2009), and particularly in cancer (Ciriello et al., 2013), copy number changes (large insertions or deletions, deleting or duplicating one or more genes). Larger scale chromosomal changes also play a role in cancer, where genome instability is common (Stephens et al., 2011). In monogenic disease, and in contrast to complex trait disease (Gusev et al., 2014; Maurano et al., 2012; Nicolae et al., 2010), very few mutations affecting expression have been identified (Landrum et al., 2016). Data for non-coding contributions in cancer are only now becoming available. Some clear examples have been identified (Horn et al., 2013; Huang et al., 2013), but a clear picture has yet to emerge. In monogenic disease, the most common mutation type is missense (Stenson et al., 2003), a single base change causing an amino acid substitution. In cancer, missense mutations also play a major role (Ciriello et al., 2013).

3.1.2 Missense mutations

In this paper I use computational methods to analyze and compare the role of missense mutations in monogenic disease and cancer. There are three primary motivations. First, as noted above, this class of mutation is the most in common in both types of disease, so that a thorough understanding of its role is worthwhile. Second, unlike most indels and copy number variants which have a major impact on protein function and hence disease phenotype, missense mutations range from no effect on protein function to complete loss of activity. The wide range of possible molecular impact presents problems for clinical interpretation. As a result, at present, evidence from computational analysis is given low weight in clinical diagnosis in monogenic disease (Richards et al., 2015). Greater understanding of how these mutations influence disease phenotype will help improve the usefulness of the computational methods. Third, with many instances of these mutations now known in both types of disease, it is possible to perform statistical analyses that provide insight into the molecular mechanisms involved.

3.1.3 Methods for interpretation of missense mutations

Methods for imputing the disease relevance of missense mutations fall into two classes: Those that rely on the pattern of observed amino acid substitutions at a mutation position both across species and paralogs and as common variants within the human population, and those that make use of structural and other molecular function information. Sequence-based methods usually utilize machine learning, and typical features are related to sequence conservation and the pattern of substitutions at the

position of interest (Cooper & Shendure, 2011). An advantage of these methods is that, provided there is a deep enough, diverse enough, and stable alignment, any mutation can be analyzed (currently 92% of the reference set of monogenic disease mutations (Stenson et al., 2003) using SNPs3D profile (Yue & Moult, 2006)). Further, subject to the assumption below, they are effective for all types of underlying mechanisms including gain of function (highly relevant for mutations in oncogenes). The disadvantage is that they provide no insight into the mechanism by which a mutation is involved in disease. The methods implicitly assume that if a mutation plays a causative role in disease, it will affect Darwinian fitness, and thus tend to be selected against. Since many monogenic diseases are early onset and severe enough to affect reproductive success, that may be a reasonable assumption. Relevance to cancer, where driver mutations promote cell growth in many ways, is less obvious. Although the assumption of a relationship between an effect on fitness and disease phenotype is embedded in the methods, there is no formal theoretical framework for calculating fitness impact. Rather machine learning (Adzhubei et al., 2010; Carter et al., 2013; Douville et al., 2016; Kircher et al., 2014; Yue & Moult, 2006) or other parameterization (Lichtarge et al., 1996; Ng & Henikoff, 2003) is used to calibrate the relationship between amino acid substitution patterns and disease phenotypes in an *ad hoc* way.

Protein structure and function provide a complementary, more mechanism oriented approach to identifying disease-relevant mutations. Previous studies have shown that a high fraction of monogenic disease and to some extent cancer tumor suppressors

mutations destabilize three dimensional-structure (Shi & Moult, 2011; Z. Wang & Moult, 2001; Yue et al., 2005): for a reference set of monogenic disease proteins (Stenson et al., 2003), SNPs3D_Stablity (Yue et al., 2005) assigned 72% as destabilizing, and for the cancer set (Ciriello et al., 2013), 50%~60% (Shi & Moult, 2011). Thus, methods of estimating the change in free energy difference between the folded and unfolded states introduced by an amino acid substitution play an important role. Molecular dynamics free energy perturbation methods (Seeliger & de Groot, 2010) may be used for this purpose. Up to now, these methods have found limited application in studies of mutations because of relatively high computational cost and lower accuracy when compared with more empirical approaches. In this paper I use SNPs3D_Stability (Yue et al., 2005) to examine the role of destabilization in both monogenic disease and cancer. The method uses empirical potential terms representing van der Waals interactions, electrostatics, conformational strain, solvent accessibility and local flexibility in a non-linear support vector machine model and was trained using monogenic disease data (Stenson et al., 2003) together with interspecies variants as controls. It has been benchmarked against experimental $\Delta\Delta G$ data and monogenic disease mutations (Yue et al., 2005). The method returns a binary yes/no estimate of whether a missense variant destabilizes a structure sufficiently to contribute to monogenic disease, together with score related to the confidence of the assignment.

3.1.4 Identifying driver mutations

There are well-established databases of causative mutations for monogenic disease (Stenson et al., 2003), and although these sources are not error-free (Xue et al., 2012), they are sufficiently accurate for many statistical purposes. Reliable identification of cancer driver mutations remains a major problem, because of the high background of passenger mutations. Current strategies focus on first identifying a subset of genes that contain driver mutations ('driver genes') and then determining which mutations in those genes are drivers. Driver genes are identified on the basis of containing a statistically higher number of cancer somatic mutations than sample background, together with other factors (Davoli et al., 2013; Dees et al., 2012; Gonzalez-Perez & Lopez-Bigas, 2012; Leiserson et al., 2015; Mermel et al., 2011; Reimand & Bader, 2013; Rubio-Perez et al., 2015; Tamborero, Gonzalez-Perez, Perez-Llamas, et al., 2013; Tamborero, Gonzalez-Perez, & Lopez-Bigas, 2013; Vogelstein et al., 2013). Although a number of sets of driver genes have been proposed, there is limited agreement between them (Tokheim, Papadopoulos, Kinzler, Vogelstein, & Karchin, 2016). It is very likely that some other genes contain some driver mutations, and conversely it is clear that driver genes will contain some level of non-driver ('passenger') mutations. In this work I used the driver gene sets derived from the two cancer mutation datasets I analyzed (Ciriello et al., 2013; Futreal et al., 2004).

A number of methods for identifying individual driver missense mutations (Carter et al., 2009; Gonzalez-Perez et al., 2012; J. S. Kaminker et al., 2007; Mao et al., 2013; Reva et al., 2011; Shihab et al., 2013) have been developed. These primarily utilize

combinations driver gene lists, predicted impact of mutations, clustering of mutations, and the number of samples in which a mutation has been observed. A limited number of driver mutations have been reliably annotated, for example, a set of 889 (Catalog of Validated Oncogenic Mutations, https://www.cancergenomeinterpreter.org/mutations). Currently, these sets are too small for a statistical analysis of properties and because of the way they were derived (an emphasis on repeat occurrence for instance) likely have significant biases. I address the problem of uncertain driver mutation assignments by considering all mutations found in the sets of driver genes, and investigating properties of interest as a function of driver assignment confidence.

3.1.5 Questions addressed

We use the sets of monogenic disease and cancer driver mutations together with the computational methods to address the following questions:

How effective are sequence-based methods for identifying mutations relevant to the two types of disease? As noted above, these methods depend on mutations impacting Darwinian fitness and, especially for cancer, the validity of that assumption is not clear. Technical issues may also limit accuracy.

How important are intrinsically disordered regions of proteins compared with ordered regions in the two types of disease? The role of disordered regions in protein function has been much discussed (Midic, Oldfield, Dunker, Obradovic, & Uversky, 2009; Vacic et al., 2012) and disease mutation properties provide insight into that.

What is the relative role of mutations on the protein surface versus those in the core of protein structures? Surface mutations are more likely to be involved in inter-molecular interactions and other mechanisms, while core mutations will be enriched for effects on protein structure stability.

How extensive is the role of destabilization of protein structure in the two types of disease? As noted above, this mechanism plays a major role in monogenic disease, but its role in cancer has been less clear.

What are the properties of mutations in cancer passenger genes? Are these benign, as the 'passenger' designation implies?

3.2 Methods

3.2.1 Monogenic disease data and cancer data

The monogenic disease set comprises 10,865 disease-related variants collected from an earlier version of HGMD (Stenson et al., 2003), together with 13,499 interspecies variants in these genes, compiled by comparing mammalian homolog protein sequences with at least 90% sequence identity over at least 80% of the full length and excluding any known disease-related variants (Yue & Moult, 2006). The disease genes can be classified as dominant or recessive based on their inheritance patterns.

Two cancer driver data sets were compiled as follows. One set was extracted from the level 3 TCGA (Cancer Genome Atlas Network, Bainbridge, et al., 2012; Cancer Genome Atlas Network, Koboldt, et al., 2012; Cancer Genome Atlas Research

Network et al., 2008, 2011, 2012; Cancer Genome Atlas Research Network, Getz, et al., 2013; Cancer Genome Atlas Research Network, Ley, et al., 2013) data described in (Ciriello et al., 2013), which has 449,788 unique somatic single-residue substitutions in a total of 3,477 tumor samples from studies on 12 different cancer types (Table 3-1). The TCGA driver set consists of the 9,325 unique somatic missense mutations found in 193 driver genes identified by (Ciriello et al., 2013). Another 415,090 somatic missense mutations in genes not belonging to the 193 driver gene list were extracted to form the TCGA passenger set. Mutations in other potential driver genes (Kumar, Searleman, Swamidass, Griffith, & Bose, 2015; Lawrence et al., 2014; Martincorena & Campbell, 2015; Tokheim et al., 2016; Vogelstein et al., 2013) were also omitted in the passenger set. 27 oncogenes and 47 tumor suppressor genes were identified in the TCGA 193 driver gene list, based on the literature (Kumar et al., 2015; Tokheim et al., 2016; Vogelstein et al., 2013), providing a TCGA Oncogene set of 1,362 missense mutations and TCGA TSG set of 2,933 missense mutations. A set of 3,116 interspecies variants in the 193 TCGA driver genes were extracted using the same procedure as for the monogenic disease set described above.

The second cancer data set was extracted from the 531,728 unique somatic single-residue substitutions in the COSMIC Database (Forbes et al., 2017) version 68. The Cosmic Gene Census (CGC) driver dataset consists of the 30,773 missense mutations in 477 driver genes identified by the Cancer Gene Census (Futreal et al., 2004). Another 495,530 missense mutations extracted from the COSMIC non-CGC genes form the CGC passenger set. Mutations in other potential driver genes were removed

in the same way as for the TCGA passenger set. A CGC Oncogene set of 7,422 missense mutations in 79 genes and a CGC TSG set of 12,016 missense mutations in 81 genes were compiled using the same procedure as for the TCGA sets. A CGC interspecies variants set of 6448 missense mutations was extracted using the same procedures as above.

Table 3-1. TCGA data set

Tumor type	TCGA ID	Number of samples	Number of unique mutations[a] in driver genes
Bladder urothelial carcinoma	BLCA	100	17431
Breast invasive carcinoma[b]	BRCA	513	17460
Colon and rectum adenocarcinoma[c]	COADREAD	498	100020
Glioblastoma multiformae[d]	GBM	276	21531
Head and neck squamous cell carcinoma	HNSC	306	34079
Kidney renal clear-cell carcinoma	KIRC	473	15557
Acute myeloid leukemia[e]	LAML	201	2500
Lung adenocarcinoma	LUAD	230	44092
Lung squamous cell carcinoma[f]	LUSC	177	44883
Ovarian serous cystadenocarcinoma[g]	OV	456	17819
Uterine corpus endometrioid carcinoma[h]	UCEC	247	106141

[a]Restricted to somatic nonsynonymous single-residue substitutions observed in samples of each specific cancer types.
[b]Reference see (Cancer Genome Atlas Network, Koboldt, et al., 2012)

[c]Reference see (Cancer Genome Atlas Network, Bainbridge, et al., 2012)
[d]Reference see (Cancer Genome Atlas Research Network et al., 2008)
[e]Reference see (Cancer Genome Atlas Research Network, Ley, et al., 2013)
[f]Reference see (Cancer Genome Atlas Research Network et al., 2012)
[g]Reference see (Cancer Genome Atlas Research Network et al., 2011)
[h]Reference see (Cancer Genome Atlas Research Network, Getz, et al., 2013)

3.2.2 Missense mutation analysis methods

Seven sequence-based missense analysis methods were used to assign missense mutations as deleterious or benign and the fraction of those mutations that are assigned as deleterious (the PDF, predicted deleterious fraction) was then calculated. Four of these (SNPs3D Profile (Yue & Moult, 2006), PolyPhen-2 (Adzhubei et al., 2010), CADD (Kircher et al., 2014), VEST3 (Carter et al., 2013; Douville et al., 2016)) were trained on monogenic disease mutation datasets (except CADD which was trained differently). The rest three: SIFT (Ng & Henikoff, 2003), LRT (Chun & Fay, 2009), and PROVEAN (Choi et al., 2012) rely on direct measures of sequence conservation properties and do not require training. In addition, three sequence methods trained specifically for interpreting cancer mutations were tested: FATHMM (Shihab, Gough, Cooper, Day, & Gaunt, 2013), Mutation Assessor (Reva et al., 2011) and CHASM (Carter et al., 2009; Wong et al., 2011). SNPs3D Profile results were generated using standalone in-house software. The dbNSFP2.9 database (X. Liu et al., 2013) was used to obtain PolyPhen-2, CADD, SIFT, LRT, VEST3, PROVEAN, FATHMM and Mutation Assessor results. CHASM results were obtained from the CRAVAT Web server (http://www.cravat.us/CRAVAT/). Binary assignments of structure destabilizing/non-destabilizing were obtained using SNPs3D Stability (Yue,

Li, & Moult, 2005) with in-house software and the predicted destabilizing fractions (PDFs) were calculated from those data.

Binary predictions were collected for PolyPhen-2, SIFT, LRT, PROVEAN, FATHMM and Mutation Assessor. The HumDiv version of PolyPhen-2 was used, and "probably damaging" and "possibly damaging" predictions were considered deleterious. MutationAssessor "H" and "M" predictions were also considered deleterious. Three methods (CADD, VEST3, and CHASM) reported continuous scores rather than binary assignments. Dataset-specific score thresholds were chosen for these, such that the false positive rates on the corresponding interspecies variants sets are similar to that of other methods. For the monogenic disease data, the score thresholds are 22 for CADD, 0.5545 for VEST3, and 0.095 for CHASM. On the TCGA data, the thresholds are 21.35 for CADD, 0.2815 for VEST3, and 0.1395 for CHASM. On the Cosmic data, the thresholds are 21.35 for CADD, 0.2915 for VEST3, and 0.1225 for CHASM.

To assess potential training bias, SNPs3D Profile and SNPs3D Stability methods were retrained on the two cancer data sets and on specific subsets of monogenic disease data. In retraining, all parameters in the support vector machine (SVM) models were re-optimized with a grid search algorithm. At each search step, the corresponding data set was bootstrapped 30 times, with the model trained on a set of randomly drawn data (number of data points equal to the data set size), and evaluated

on the data points not included in training. 95% confidence intervals in the other analyses were also inferred from 30 rounds of bootstrapping.

3.2.3 Structure modeling

For analysis of structure-related features, the set of experimental protein structures was extended by building homology models for protein domains where a suitable template was available, as described in (Yue et al., 2005). The procedure is briefly summarized here. Proteins that have > 40% sequence identity to the query protein and a crystal structure of < 3Å resolution are used as templates for backbone conformations. The 40% sequence identity cutoff is based on earlier benchmarking (Yue et al., 2005) that showed prediction accuracy for models based on 40% or higher sequence identity to a template is not significantly lower than for that based on experimental structures. Where the template amino acids are identical to the corresponding ones in the query structure, side chains atoms from the template are used. Otherwise, the side-chains are modeled using SCWRL (Canutescu, Shelenkov, & Dunbrack, 2003).

3.2.4 Analysis of somatic missense mutation recurrence and density

For each unique cancer somatic missense mutation, the recurrence was calculated as the number of times the mutation was observed in all samples. The majority of unique somatic missense mutations, even in the likely driver genes, have a low recurrence (< 5). Recurrence values were grouped into six bins, the last of which covers all recurrence larger than 10. The cancer type specific mutation load was defined as the

average number of unique missense mutations per sample, observed in the samples of the corresponding cancer type.

3.2.5 Analysis of structure disorder and surface missense mutations

DISOPRED3.16 (Jones & Cozzetto, 2015) with default parameters was used to predict intrinsically disordered protein residues in each data set. STRIDE (Eisenhaber & Argos, 1993; Eisenhaber et al., 1995; Frishman & Argos, 1995) was used to calculate the absolute solvent accessible surface area (SASA) of each amino acid residue in the protein structures or homology models prepared as described in (Yue et al., 2005) and above. The relative SASA was then calculated by normalizing the STRIDE results with the corresponding amino acid residue maximal solvent accessibility reported in (Rost & Sander, 1994). Based on the relative SASA, residue location was assigned as buried core (<0.05), partially exposed (≥0.05, ≤ 0.25), and surface (>0.25). The relative density (*RD*) of missense mutations in a particular state is calculated as follows:

$$RD = \frac{N_i/M_i}{N_j/M_j}$$

where, given two particular states i and j (disordered, ordered, buried in the core, and exposed on the surface), N_i and N_j are the total number of missense mutations in the corresponding states, and M_i and M_j are the total number of amino acid residues in the corresponding states.

3.3 Results

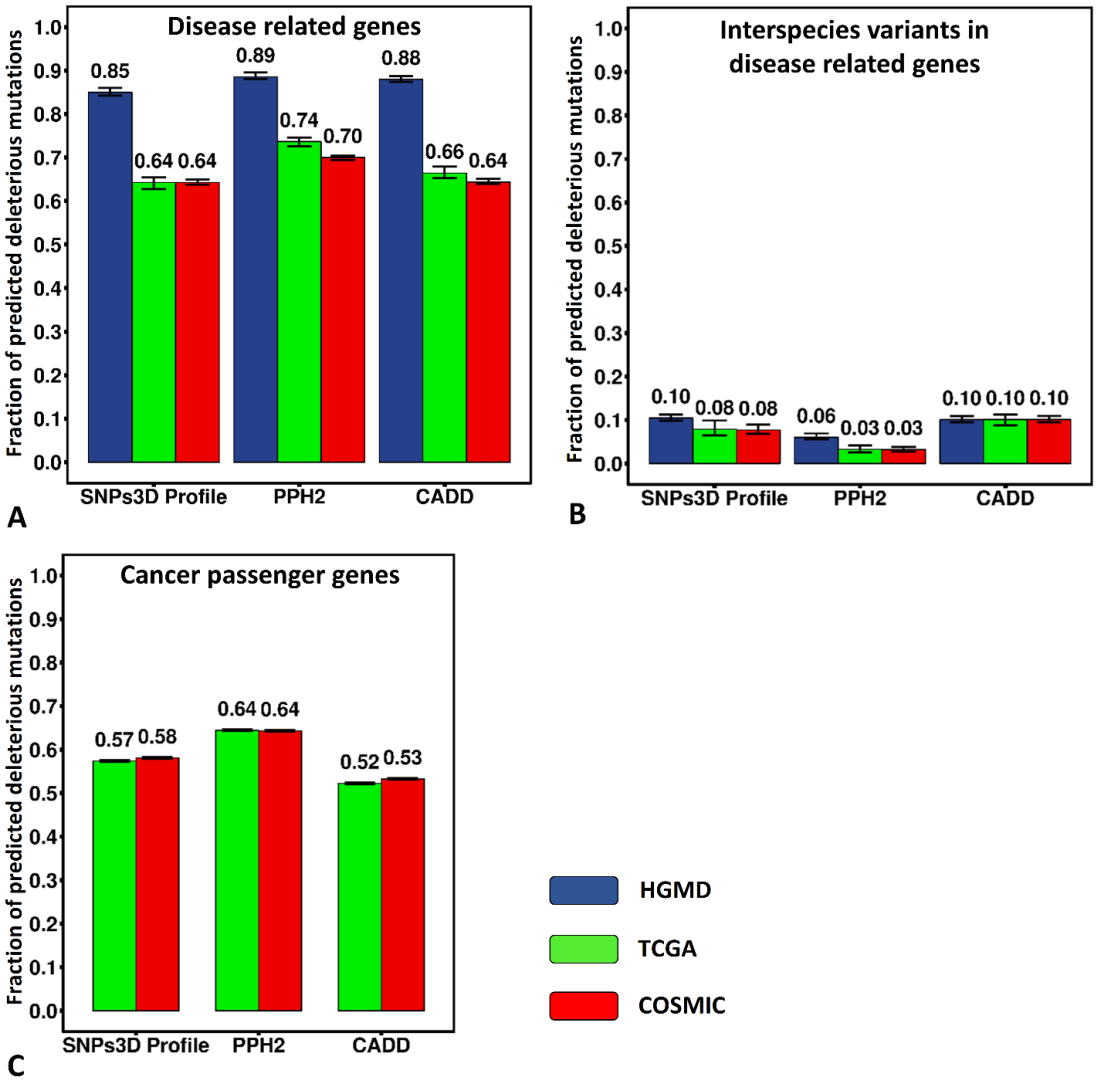
3.3.1 Performance of variant interpretation methods on monogenic disease and cancer missense mutations

I begin the analysis by investigating the fraction of monogenic disease mutations and assumed cancer drivers that are predicted to be deleterious by a number of sequence-based methods. As noted earlier, these methods indirectly utilize the impact of a mutation on fitness. Figure 3-1A shows the results using three different sequence methods (SNPs3D profile (Yue & Moult, 2006), Polyphen2 (Adzhubei et al., 2010) and CADD (Kircher et al., 2014)), comparing the fraction of mutations predicted to be deleterious on a monogenic disease dataset, HGMD (Stenson et al., 2003) and mutations in two sets of cancer driver genes (Cancer Genome Atlas Network, Bainbridge, et al., 2012; Cancer Genome Atlas Network, Koboldt, et al., 2012; Cancer Genome Atlas Research Network et al., 2008, 2011, 2012; Cancer Genome Atlas Research Network, Getz, et al., 2013; Cancer Genome Atlas Research Network, Ley, et al., 2013; Forbes et al., 2017). These methods have previously been shown to be effective for identifying monogenic disease mutations (Dong et al., 2015; Yin, Kundu, Pal, & Moult, 2017; Yue et al., 2005), and consistent with that, the fraction predicted deleterious here is high, between 0.85 and 0.88. For all three methods, the fraction deleterious for mutations in cancer driver genes is substantially lower (0.64 - 0.74) (Figure 3-1A). An additional four monogenic disease missense analysis methods and three developed specifically for cancer missense analysis show the same pattern (Table 3-2). Previous studies have also shown similar results for cancer data (Gnad et al., 2013; Martelotto et al., 2014).

Figure 3-1B shows that in contrast to the results for the disease mutations, interspecies variants in these three sets of genes have a uniformly low predicted deleterious fraction (0.03 - 0.10), with no significant differences between monogenic disease and cancer. Although a low number of human monogenic disease mutations may be found to be fixed in other species (Kondrashov, Sunyaev, & Kondrashov, 2002), these numbers do provide an approximate measure of the false positive rate for the methods. Surprisingly, for mutations in passenger genes (Figure 3-1C) the predicted deleterious fraction is much higher (34% to 82%) than for the interspecies variants in the driver genes. As discussed later, there are two possible factors contributing here: weak purifying selection in cancer samples (suggested by others (McFarland, Korolev, Kryukov, Sunyaev, & Mirny, 2013)), and additional drivers in genes currently designated as passengers. Table 3-3 provides the number of missense mutations and other statistics for all datasets.

Figure 3-1: Performance of three sequence-based variant interpretation methods on mutations in the two types of diseases. (**A**) Fraction of predicted deleterious mutations for monogenic disease mutations (HGMD, blue), cancer somatic mutations in the Sander driver gene list (TCGA Sander, green), and cancer somatic mutations in the COSMIC Cancer Gene Census driver gene list (COSMIC, red). A consistently high fraction of monogenic disease mutations appear deleterious, while the fractions of mutations in cancer driver genes are consistently lower. (**B**) Fraction of predicted deleterious interspecies variants in these gene sets. A very low fraction is predicted deleterious in all three data sets, supporting a low false positive rate for the methods. (**C**) Fraction of predicted deleterious mutations in passenger genes in the two cancer data sets. A surprisingly high fraction of mutations are predicted deleterious compared with the interspecies controls, suggesting limited purifying selection and the presence of additional driver mutations.

Error bars show 95% confidence intervals derived from 100 rounds of bootstrapping.



(Figure 3-1. See above for caption.)

Table 3-2. Performance of sequence-based variant interpretation methods on all datasets

	Variant interpretation method	Fraction of predicted deleterious mutations			Fraction of predicted benign interspecies variants			Fraction of deleterious cancer missense mutations	
		HGMD	TCGA Sander	Cosmic Gene Census	HGMD	TCGA Sander	Cosmic Gene Census	TCGA Passenger	Cosmic Passenger
General	SNPs3D Profile	0.85	0.64	0.64	0.90	0.92	0.92	0.43	0.42
	PPI2	0.89	0.74	0.70	0.94	0.97	0.97	0.36	0.36
	CADD	0.88	0.66	0.64	0.90	0.90	0.90	0.48	0.47
	SIFT	0.76	0.67	0.65	0.90	0.99	0.99	0.43	0.42
	LRT	0.83	0.73	0.72	0.89	1.00	1.00	0.42	0.41
	VEST3	0.92	0.81	0.79	0.90	0.99	0.90	0.35	0.34
	PROVEAN	0.81	0.55	0.53	0.92	0.96	0.97	0.55	0.54
	SNPs3D Stability	0.72	0.49	0.47	0.87	0.86	0.88	0.59	0.58
Cancer specific	FATHMM	0.81	0.38	0.40	0.39	0.68	0.70	0.82	0.81
	Mutation Assessor	0.74	0.48	0.45	0.95	0.97	0.97	0.60	0.58
	CHASM	0.43	0.49	0.61	0.90	0.90	0.90	0.76	0.78

Table 3-3. Total number of mutations and genes (Italics in brackets) in each dataset, and coverage of these by different variant interpretation methods, for each dataset used.

Prediction Method	HGMD	HGMD Interspecies	TCGA Sander	TCGA Sander Interspecies	TCGA Passenger	TCGA Sander Oncogene	TCGA Sander Tumor Suppressor	Cosmic Gene Census	Cosmic Gene Census Interspecies	Cosmic Gene Passenger	Cosmic Gene Census Oncogene	Cosmic Gene Census Tumor Suppressor
SNPs3D Profile	10004 (*745*)	10121 (*284*)	6706 (*203*)	1834 (*74*)	307829 (*17185*)	1131 (*28*)	1970 (*50*)	22763 (*546*)	3492 (*180*)	380489 (*15860*)	6215 (*87*)	7859 (*103*)
PPH2	10002 (*727*)	10415 (*266*)	8412 (*212*)	2659 (*77*)	401899 (*17544*)	1354 (*29*)	2886 (*58*)	29540 (*519*)	5558 (*187*)	494897 (*16098*)	7123 (*86*)	11544 (*98*)
CADD	9869 (*724*)	10351 (*267*)	8454 (*214*)	2678 (*77*)	404224 (*17621*)	1354 (*29*)	2887 (*58*)	29559 (*520*)	5586 (*187*)	495491 (*16139*)	7123 (*86*)	11545 (*98*)
SIFT	4235 (*502*)	6855 (*237*)	8379 (*214*)	2678 (*77*)	400115 (*17433*)	1354 (*29*)	2838 (*58*)	29276 (*516*)	5556 (*185*)	493072 (*16024*)	7107 (*84*)	11287 (*97*)
LRT	9647 (*699*)	10028 (*262*)	7682 (*174*)	2527 (*76*)	359929 (*16210*)	1285 (*25*)	2783 (*48*)	28114 (*507*)	5469 (*186*)	455807 (*15292*)	6807 (*84*)	10772 (*93*)
VEST3	10004 (*723*)	10438 (*265*)	8405 (*210*)	2668 (*77*)	402427 (*17596*)	1354 (*29*)	2885 (*58*)	29554 (*520*)	5558 (*187*)	495491 (*16139*)	7123 (*86*)	11544 (*98*)
PROVEAN	10004 (*723*)	10438 (*265*)	8426 (*213*)	2668 (*77*)	402391 (*17590*)	1354 (*29*)	2887 (*58*)	29559 (*520*)	5558 (*187*)	495322 (*16138*)	7123 (*86*)	11545 (*98*)
SNPs3D Stability	6096 (*460*)	1771 (*194*)	3067 (*133*)	297 (*46*)	79764 (*6810*)	765 (*28*)	1264 (*41*)	11431 (*359*)	447 (*89*)	100579 (*6448*)	4058 (*73*)	4819 (*80*)
FATHMM	665 (*233*)	549 (*106*)	8361 (*211*)	2643 (*75*)	383121 (*16183*)	1350 (*29*)	2876 (*57*)	29148 (*507*)	5333 (*181*)	473737 (*14960*)	7059 (*86*)	11543 (*97*)
Mutation Assessor	657 (*232*)	547 (*105*)	8375 (*200*)	2658 (*77*)	398147 (*16957*)	1349 (*28*)	2860 (*54*)	29311 (*513*)	5523 (*186*)	490181 (*15705*)	7080 (*83*)	11449 (*98*)
CHASM	10865 (*784*)	13484 (*288*)	9325 (*234*)	3116 (*82*)	412778 (*17698*)	1362 (*30*)	2933 (*61*)	30767 (*554*)	6448 (*197*)	494101 (*16063*)	7421 (*90*)	12016 (*104*)
DISOPRED3	10865 (*784*)	13499 (*288*)	9233 (*233*)	3116 (*82*)	410792 (*17715*)	1362 (*30*)	2933 (*61*)	30773 (*554*)	6448 (*197*)	493231 (*16061*)	7422 (*90*)	12016 (*104*)
[a]Ordered	9755 (*738*)	9371 (*282*)	6379 (*215*)	1474 (*71*)	296136 (*17414*)	1040 (*29*)	1935 (*52*)	20210 (*537*)	2339 (*175*)	362346 (*15924*)	5927 (*88*)	7563 (*102*)
[a]Disordered	1110 (*261*)	4128 (*243*)	2854 (*196*)	1642 (*77*)	114656 (*13142*)	322 (*21*)	998 (*54*)	10563 (*490*)	4109 (*182*)	130885 (*12459*)	1495 (*79*)	4453 (*101*)
STRIDE	6095 (*460*)	1722 (*193*)	3055 (*132*)	297 (*46*)	79704 (*6800*)	765 (*28*)	1263 (*41*)	11428 (*359*)	447 (*89*)	100504 (*6441*)	4057 (*73*)	4818 (*80*)
[b]Core	2156 (*302*)	199 (*96*)	590 (*93*)	30 (*17*)	13568 (*4354*)	122 (*22*)	309 (*33*)	1272 (*252*)	39 (*24*)	17799 (*4391*)	643 (*59*)	1112 (*65*)
[b]Surface	2025 (*352*)	1244 (*187*)	1676 (*127*)	219 (*45*)	47645 (*6630*)	435 (*26*)	586 (*41*)	6209 (*350*)	346 (*86*)	59308 (*6311*)	2283 (*73*)	2330 (*78*)

[a]Ordered or disordered mutation subsets predicted by DISOPRED3

[b]Core or surface mutation subsets predicted by STRIDE

3.3.2 The effect of passenger mutations in cancer driver genes

A likely reason for the lower fraction of mutations predicted deleterious for cancer is that not all mutations in driver genes are drivers, for example, the mutations within the C-terminus in the APC protein (Vogelstein et al., 2013). I used two methods to identify subsets of mutations enriched in drivers. The first assumes that the more cancer samples a mutation is observed in, the more likely it is to be a driver, an approach that others have also used (Ciriello et al., 2013). Figure 3-2 shows the dependence of the predicted deleterious fraction (PDF) on the number of occurrences of a mutation. Strikingly, the PDF in driver genes increases sharply with mutation recurrence, from approximately 0.6 for mutations only observed once to 0.9 for those observed more than 10 times. The latter value is higher than the PDF for monogenic disease. For mutations in passenger genes, on the other hand, the PDF does not increase with recurrence and is lower than the lowest value for driver genes. Thus, by this criterion, the low PDF observed in the driver gene mutation set is primarily a consequence of the presence of passenger mutations in driver genes, and a pure driver set would have a PDF values as high as or higher than that for monogenic disease.

The second method of enriching for driver mutations examines the PDF as a function of the average total mutational load in different cancer types. As noted above, mutational load differs by more than two orders of magnitude, depending on cancer type (Martincorena & Campbell, 2015; Vogelstein et al., 2013), so that the background of passengers in low mutation load cancers will be very much smaller than for high load ones. Thus, if passengers in driver genes are a cause of the low

overall PDF, the PDF will be higher in cancer types with a lower mutational load. Figure 3-3 shows that this is indeed the case: for driver genes, the trend is for increased PDF as mutation load decreases, whereas passenger genes show no trend. These results are consistent across three different variant interpreting methods (Figure 3-4). Thus, by this criterion too, the low PDF observed in the driver gene mutation set is primarily a consequence of the presence of passenger mutations in driver genes.

3.3.3 Other factors that may affect the fraction of driver gene mutations predicted deleterious

I explored two other possible explanations for the different deleterious rates for monogenic disease and cancer mutations. One difference between the two types of disease is that whereas most monogenic disease missense mutations overwhelmingly result in loss of protein function (Yue et al., 2005), cancer driver mutations are either loss of function (in tumor suppressors) or gain of function (in oncogenes). I divided the cancer data into these two subtypes and repeated the analysis (Table 3-4). Results vary a little by methods, but overall there is no substantial difference between the two types of driver genes, so this is not a significant factor in the monogenic disease/cancer difference. A second possible explanation is training bias - methods trained on one type of disease may not perform as well on the other. Two lines of evidence show this is also not a significant factor. First, the analysis done with the three cancer-specific methods shows a similar monogenic disease/cancer difference (Table 3-2). Second, versions of the SNPs3D Profile method retrained on the cancer datasets also produce a lower predicted deleterious fraction in cancer genes than in monogenic disease (Table 3-5 shows statistics on this and all other retraining results).

Table 3-4. Performance of variant interpretation methods on cancer oncogene and tumor suppressor gene subsets

	Variant interpreting method	Fraction of predicted deleterious mutations in oncogenes		Fraction of predicted deleterious mutations in tumor suppressor genes	
		TCGA Sander	Cosmic Gene Census	TCGA Sander	Cosmic Gene Census
General	SNPs3D Profile	0.66	0.68	0.75	0.70
	PPH2	0.72	0.72	0.77	0.73
	CADD	0.79	0.68	0.71	0.66
	SIFT	0.69	0.67	0.75	0.68
	LRT	0.87	0.82	0.73	0.75
	VEST3	0.88	0.83	0.87	0.83
	PROVEAN	0.60	0.59	0.60	0.57
	SNPs3D Stability	0.47	0.44	0.59	0.53
Cancer specific	FATHMM	0.37	0.39	0.52	0.57
	MutationAssessor	0.52	0.45	0.53	0.49
	CHASM	0.75	0.80	0.71	0.80

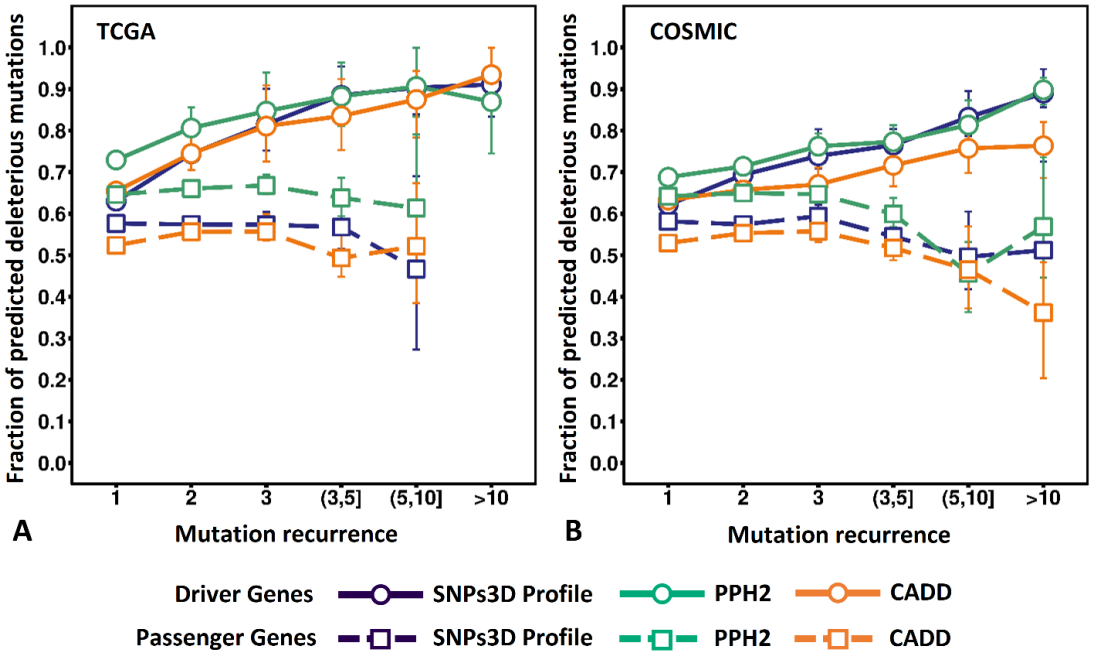


Figure 3-2: Fraction of predicted deleterious mutations in the driver genes (circles and solid lines) as a function of mutation recurrence, for two cancer datasets. The fraction rises from around 0.6 for single occurrence mutations to about 0.9 for those occurring more than 10 times. In contrast to that, for passenger gene mutations (squares and dashed lines), the value is approximately constant for all recurrence values, and lower than for the lowest recurrence driver gene value. These data are consistent with an increase in the fraction of driver mutations with recurrence. For the most enriched driver set, the fraction predicted deleterious is higher than that for monogenic disease (**Figure 3-1**). The results are consistent across SNPs3D Profile (blue), PPH2 (green) and CADD (orange). Error bars show 95% confidence intervals derived from 100 rounds of bootstrapping.

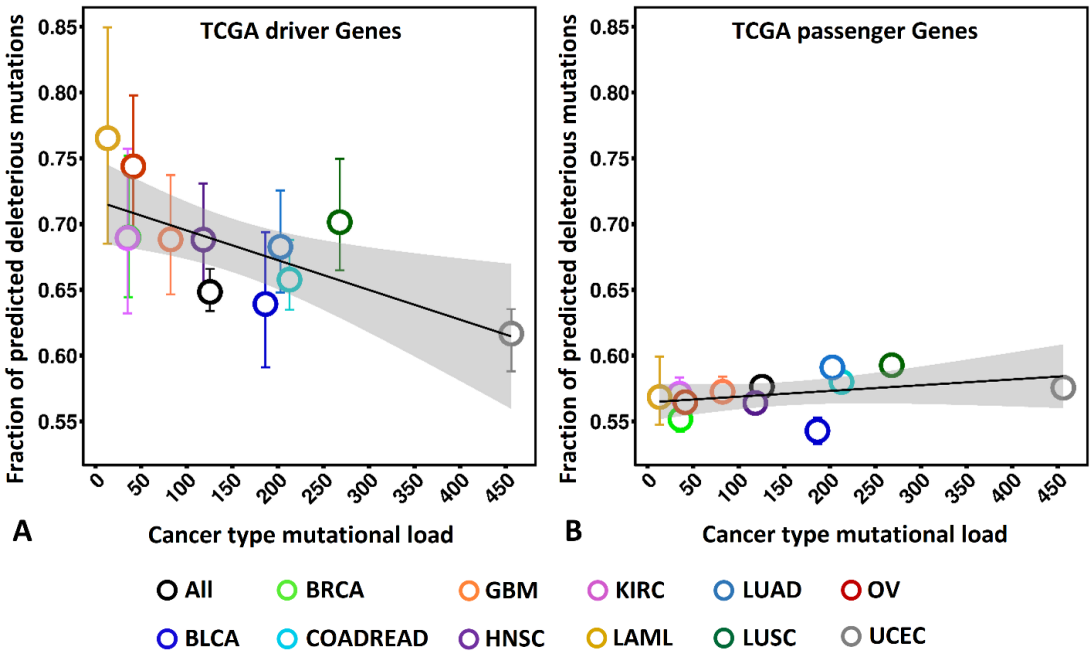
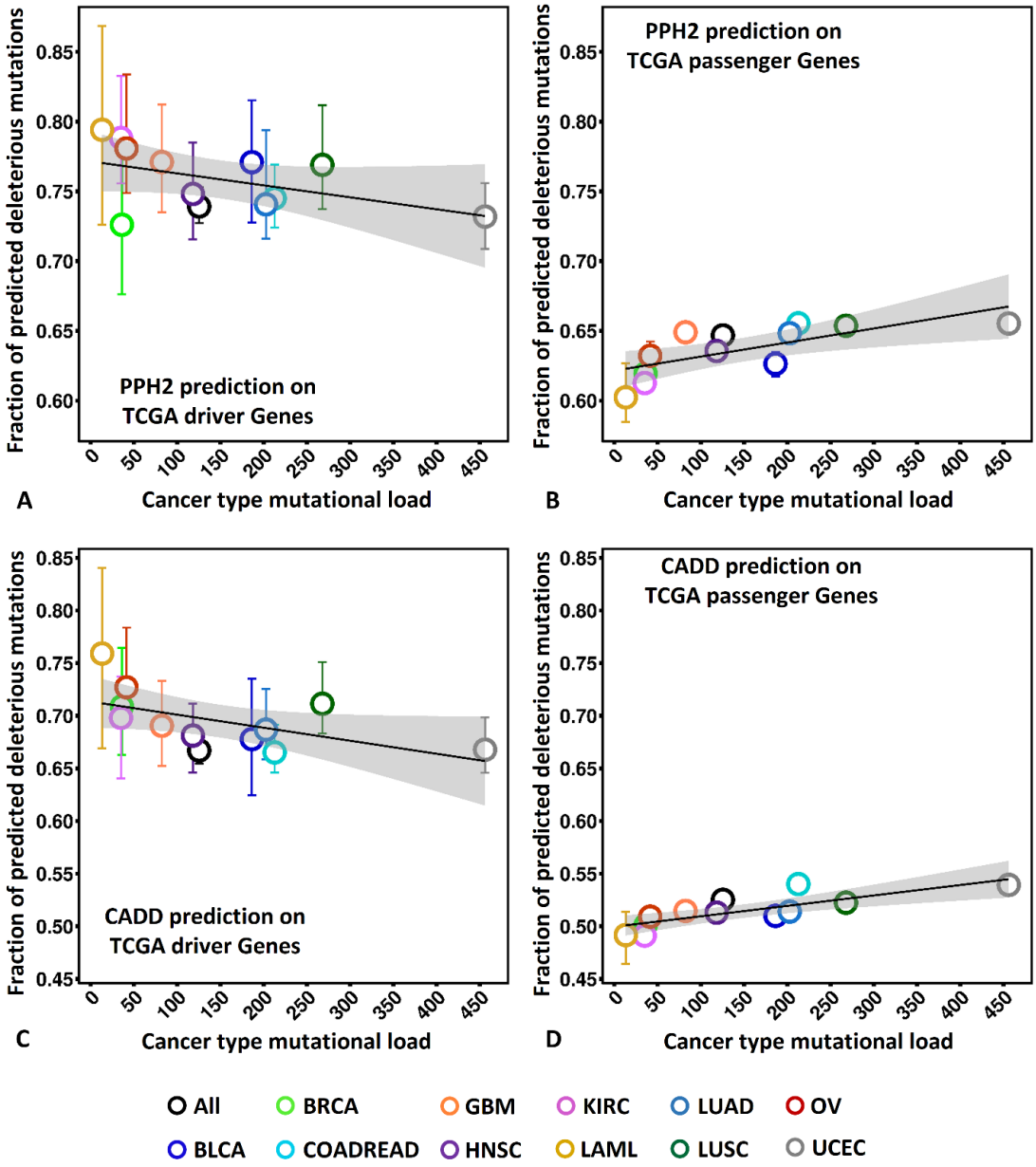


Figure 3-3: (A) The fraction of predicted deleterious mutations in TCGA Sander driver genes is negatively correlated with the mutation load across cancer types, whereas (B) the fraction of predicted deleterious mutations in the TCGA passenger genes does not show correlation with the mutation load. The correlation is consistent across SNPs3D Profile (shown here), PPH2 (shown in **Figure 3-4**) and CADD (**Figure 3-4**). The results are consistent with the overall low predicted deleterious fraction arising from the burden of passenger mutations in driver genes. Error bars show 95% confidence intervals inferred from 100 rounds of bootstrapping. Small error bars may be obscured by symbols. BLCA: Bladder urothelial carcinoma; BRCA: Breast invasive carcinoma; COADREAD: Colon and rectum adenocarcinoma; GBM: Glioblastoma multiforme; HNSC: Head and neck squamous cell carcinoma; KIRC: Kidney renal clear-cell carcinoma; LAML: Acute myeloid leukemia; LUAD: Lung adenocarcinoma; LUSC: Lung squamous cell carcinoma;

OV: Ovarian serous cystadenocarcinoma; UCEC: Uterine corpus endometrioid carcinoma.

Figure 3-4: The fraction of predicted deleterious mutations derived with (A) PPH2 and (C) CADD in driver genes in the TCGA Sander list are negatively correlated with the mutation burden across cancer types, whereas the fraction of predicted deleterious mutations in the TCGA passenger genes (**B, D**) show no or weak positive correlation with the mutation burden. These results are consistent with those from SNPs3D Profile (**Figure 3**). Error bars indicate the 95% confidence intervals inferred from 100 round bootstrapping. Small error bars may not be obscured by the symbols. BLCA, Bladder urothelial carcinoma. BLCA: Bladder urothelial carcinoma; BRCA: Breast invasive carcinoma; COADREAD: Colon and rectum adenocarcinoma; GBM: Glioblastoma multiforme; HNSC: Head and neck squamous cell carcinoma; KIRC: Kidney renal clear-cell carcinoma; LAML: Acute myeloid leukemia; LUAD: Lung adenocarcinoma; LUSC: Lung squamous cell carcinoma; OV: Ovarian serous cystadenocarcinoma; UCEC: Uterine corpus endometrioid carcinoma.



(Figure 3-4, see above for caption.)

Table 3-5. Performance of SNPs3D methods trained on specific datasets

Training Data	Testing Data	Testing subset	[a]SNPs3D Profile			[a]SNPs3D Stability		
			[b]Disease gene	Interspecies variants	Passenger gene	[b]Disease gene	Interspecies variants	Passenger gene
HGMD	HGMD	Whole	0.85	0.10	/	0.72	0.13	/
	TCGA		0.64	0.08	0.43	0.49	0.14	0.59
	COSMIC		0.64	0.08	0.42	0.47	0.12	0.58
TCGA Sander	HGMD		0.89	0.18	/	0.81	0.41	/
	TCGA		0.74	0.14	0.69	0.67	0.24	0.61
	COSMIC		0.74	0.17	0.69	0.64	0.30	0.61
[c]CGC	HGMD		0.89	0.16	/	0.84	0.39	/
	TCGA		0.72	0.13	0.65	0.68	0.32	0.62
	COSMIC		0.72	0.14	0.66	0.65	0.29	0.62
HGMD disordered and ordered	[f]HGMD	Whole	0.84	0.10	/	/	/	/
	TCGA		0.64	0.08	0.57	/	/	/
	TCGA Sander [d]OG		0.65	0.05	/	/	/	/
	TCGA Sander [e]TS		0.75	0.07	/	/	/	/
	COSMIC		0.64	0.09	0.58	/	/	/
	[c]CGC [d]OG		0.67	0.07	/	/	/	/
	[c]CGC [e]TS		0.70	0.10	/	/	/	/
HGMD ordered	HGMD	Ordered	0.85	0.11	/	/	/	/
	TCGA		0.68	0.10	0.60	/	/	/
	TCGA Sander [d]OG		0.70	0.05	/	/	/	/
	TCGA Sander [e]TS		0.78	0.08	/	/	/	/
	COSMIC		0.69	0.09	0.60	/	/	/
	[c]CGC [d]OG		0.70	0.07	/	/	/	/
	[c]CGC [e]TS		0.74	0.10	/	/	/	/
HGMD disordered	[f]HGMD	Disordered	0.69	0.06	/	/	/	/
	TCGA		0.47	0.04	0.47	/	/	/
	TCGA Sander [d]OG		0.39	0.04	/	/	/	/
	TCGA Sander [e]TS		0.62	0.04	/	/	/	/
	COSMIC		0.49	0.09	0.47	/	/	/
	[c]CGC [d]OG		0.47	0.07	/	/	/	/
	[c]CGC [e]TS		0.52	0.09	/	/	/	/
HGMD core and surface	HGMD	Whole	0.82	0.13	/	0.71	0.14	/
	HGMD Dominant		0.84	0.13	/	0.69	0.14	/
	HGMD Recessive		0.83	0.12	/	0.75	0.15	/
	TCGA		0.72	0.11	0.62	0.53	0.17	0.44
	TCGA Sander [d]OG		0.67	0.10	/	0.51	0.16	/
	TCGA Sander [e]TS		0.80	0.09	/	0.61	0.17	/
	COSMIC		0.72	0.11	0.63	0.50	0.16	0.45
	[c]CGC [d]OG		0.73	0.10	/	0.47	0.16	/
	[c]CGC [e]TS		0.74	0.06	/	0.55	0.16	/
HGMD core	HGMD	Core	0.85	0.07	/	0.76	0.09	/
	HGMD Dominant		0.86	0.13	/	0.76	0.13	/
	HGMD Recessive		0.85	0.18	/	0.79	0.11	/
	TCGA		0.74	0.10	0.67	0.65	0.17	0.54
	TCGA Sander [d]OG		0.65	0.18	/	0.59	0.18	/
	TCGA Sander [e]TS		0.83	[g]0	/	0.73	0.17	/
	COSMIC		0.76	0.05	0.68	0.61	0.10	0.56
	[c]CGC [d]OG		0.71	0.13	/	0.53	0.12	/
	[c]CGC [e]TS		0.81	[g]0	/	0.67	[f]0.25	/

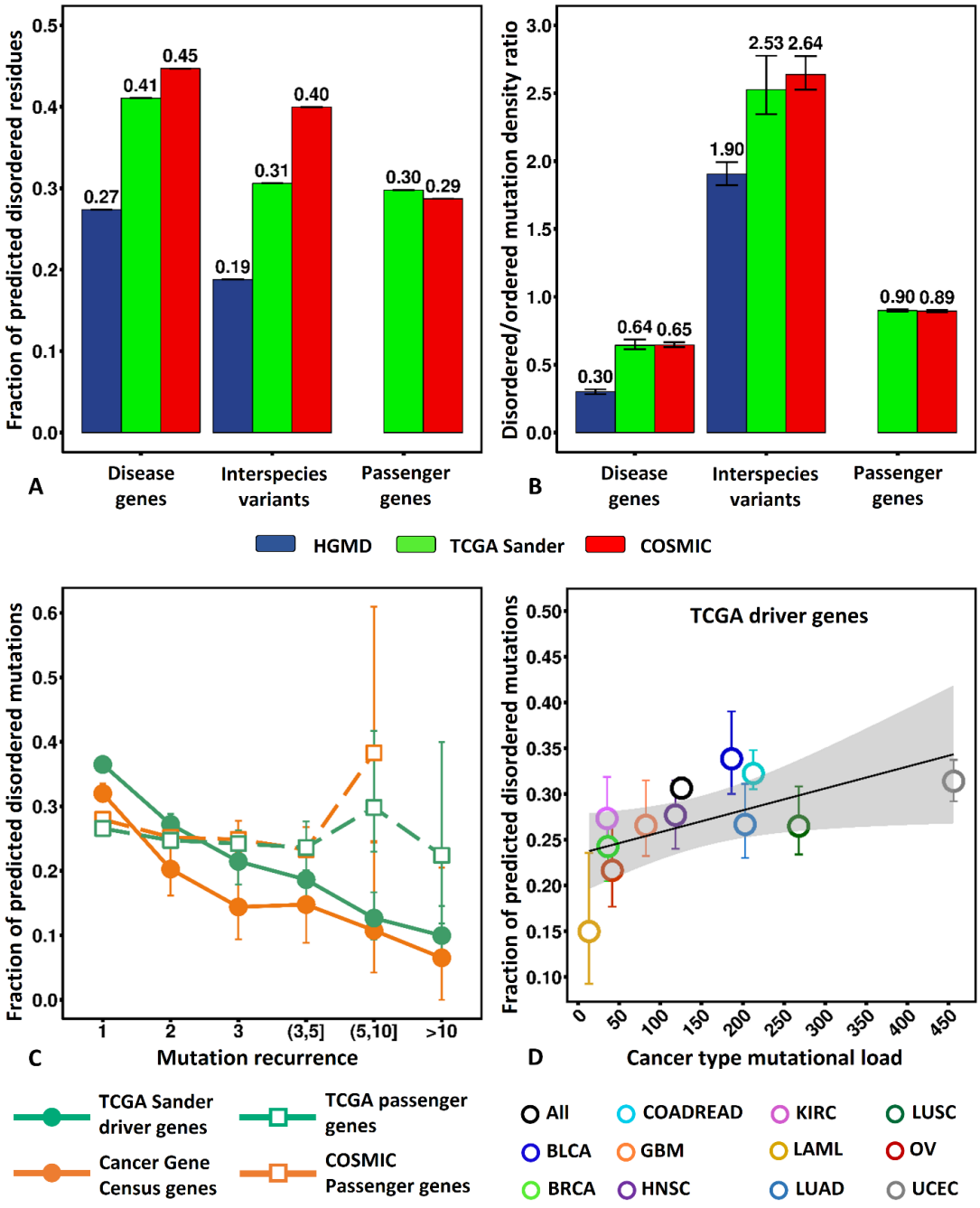
	HGMD		0.78	0.12	/	0.63	0.16	/
	HGMD Dominant		0.83	0.12	/	0.61	0.16	/
	HGMD Recessive		0.77	0.10	/	0.68	0.16	/
	TCGA		0.69	0.12	0.60	0.43	0.16	0.38
HGMD surface	TCGA Sander [d]OG	Surface	0.66	0.1	/	0.44	0.13	/
	TCGA Sander [e]TS		0.72	0.11	/	0.45	0.19	/
	COSMIC		0.69	0.12	0.60	0.42	0.16	0.38
	[c]CGC [d]OG		0.71	0.09	/	0.42	0.14	/
	[c]CGC [e]TS		0.68	0.14	/	0.45	0.15	/

[a]Fraction of missense mutations predicted deleterious or destabilizing
[b]Disease related genes in HGMD, or the Sander list of diver genes in TCGA, or Cancer Gene Census (CGC) driver genes in COSMIC
[c]CGC, Cancer Gene Census (COSMIC)
[d]OG, Oncogenes
[e]TS, Tumor suppressor genes
[f]After removing collagen proteins
[g]Low counts for this subset

3.3.4 Intrinsically disordered regions in monogenic disease and cancer

I next examine the first of three protein structure related factors that affect the properties of monogenic disease and cancer mutations: the role of intrinsically disordered structure. 27% of residues in monogenic disease proteins are predicted disordered by the method used here (Jones & Cozzetto, 2015) (Figure 3-5A). Cancer passenger proteins, representing the majority of genes, have a similar value (Figure 3-5A). But for cancer driver proteins, as others have also noted (Pajkos, Mészáros, Simon, & Dosztányi, 2012), the predicted content of disordered residues is substantially higher, at 40 to 45% (Figure 3-5A). (The disorder data are derived using a machine learning prediction method (Jones & Cozzetto, 2015) rather than direct observation of structure, 5% false positive rate threshold).

Figure 3-5: (**A**) Predicted fraction of intrinsically disordered residues. Only about ¼ of monogenic disease protein residues are predicted disordered (blue), compared with nearly twice as many in cancer driver genes (green and red). Passenger gene values are close to those for monogenic disease. (**B**) Ratio of mutation density in disordered regions to that in ordered regions. The relative density of cancer driver mutations is more than twice as high as for monogenic disease. High passenger protein relative density and very high values for interspecies variants are consistent with lower functional restraints in disordered regions. (**C**) The fraction of mutations in disordered regions of cancer driver genes decreases with mutation recurrence rate, consistent with most mutations in these regions being passengers. No dependence on recurrence rate is seen for the equivalent mutations in passenger genes. (**D**) The fraction of mutations in disordered regions of cancer driver proteins increases with cancer type mutational load, also consistent with most of these mutations being passengers. Error bars show 95% confidence intervals derived from 100 rounds of bootstrapping. Small error bars may be obscured by symbols. BLCA: Bladder urothelial carcinoma; BRCA: Breast invasive carcinoma; COADREAD: Colon and rectum adenocarcinoma; GBM: Glioblastoma multiforme; HNSC: Head and neck squamous cell carcinoma; KIRC: Kidney renal clear-cell carcinoma; LAML: Acute myeloid leukemia; LUAD: Lung adenocarcinoma; LUSC: Lung squamous cell carcinoma; OV: Ovarian serous cystadenocarcinoma; UCEC: Uterine corpus endometrioid carcinoma.



(Figure 3-5. See above for caption.)

3.3.5 Mutations in intrinsically disordered regions

Figure 3-5B shows that monogenic disease mutations are only ~1/3 as likely to occur at disordered positions as ordered ones. This, together with the low fraction of disordered residues, results in a total of only 10% of monogenic disease mutations lying in disordered regions, indicating a small role for these in this type of disease. In contrast to this, cancer driver gene mutations are only moderately less likely in disordered regions than in ordered ones (0.76), and that, together with the higher content of disorder in cancer driver proteins, results in total 31~34% of these mutations occurring in disordered regions. Two factors may contribute to the higher disorder cancer mutation density - an excess of passenger mutations in disordered regions, and a possible greater functional role for disordered regions in cancer drivers than in monogenic disease. Figure 3-5B also shows that the density of mutations in the ordered and disordered regions of passenger proteins is approximately equal and slightly higher than that of driver proteins and that the highest relative density is for interspecies variants, with a density more than 2.5 times higher in disordered regions than ordered ones. Both this and the higher passenger relative density are consistent with substantially less functional restrictions on the acceptance of mutations in disordered regions and so a tendency for passengers to accumulate there. In support of this, Figure 3-5C shows that the fraction of driver gene mutations observed more than 10 times (and therefore most likely to drivers) in disordered regions is only 1/3 the fraction for mutations observed only once (so least likely to be drivers). Similarly, Figure 3-5D shows that the fraction of mutations in disordered regions decreases with

decreasing total mutational load, consistent with a higher fraction of passengers in these regions.

3.3.6 Fraction of deleterious mutations in ordered and disordered regions

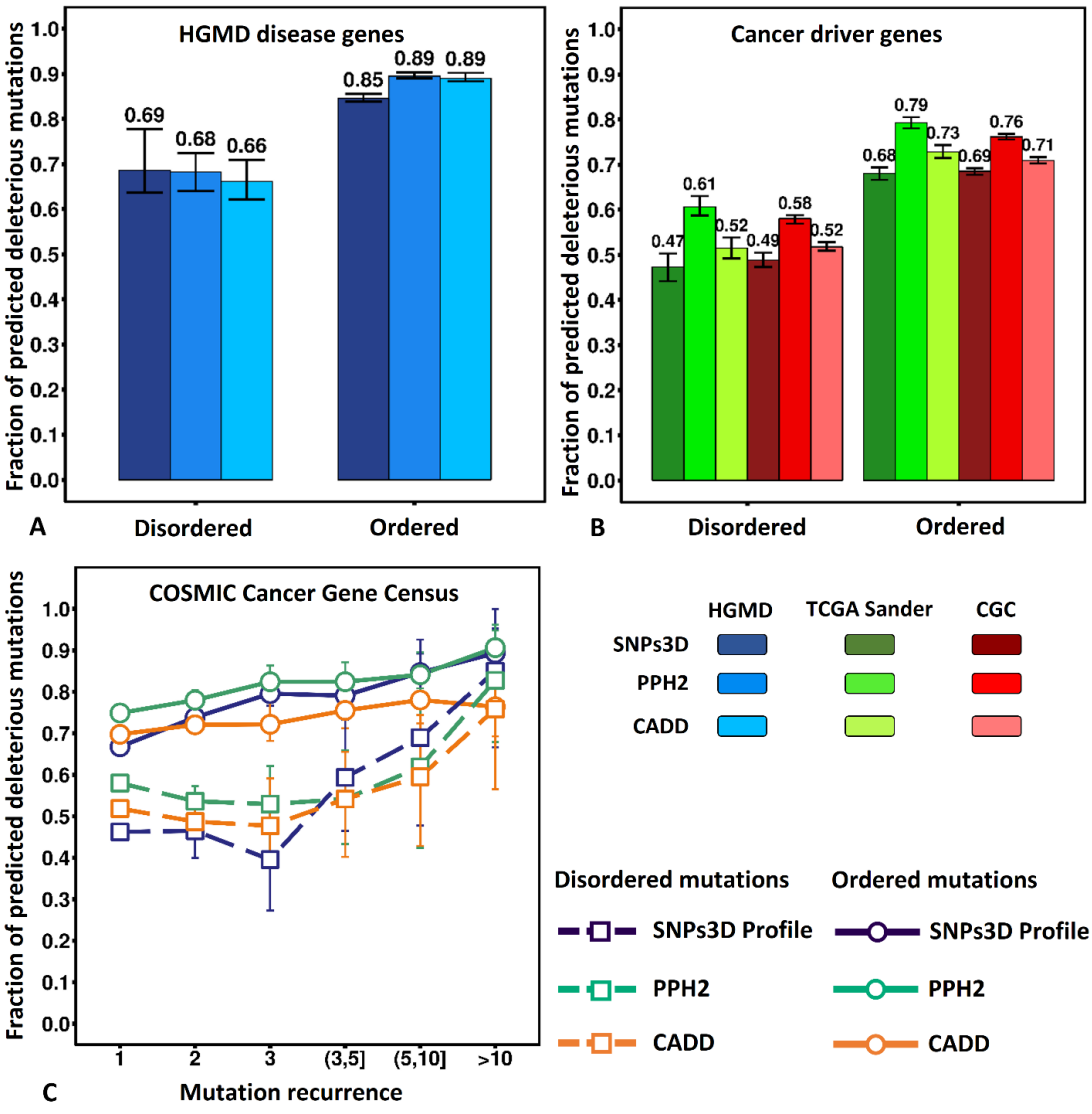
I next examine predicted deleterious rates in disordered versus ordered regions, using the sequence methods introduced earlier. Since all sequence methods are trained on a full set of disease mutations, and, particularly for monogenic disease, there are more mutations in ordered than disordered regions, training bias was a concern here. To investigate the extent of bias, I trained versions of the SNPs3D profile method on only disordered HGMD mutations together with disordered interspecies variants as controls and also trained on the corresponding ordered data. In fact, retraining made almost no difference to the results: the predicted deleterious fraction (PDF) in ordered regions is unchanged at 0.85. The original PDF in disordered regions is 0.86 and for the retrained method is 0.84. Correction for a second factor does have a significant impact on the results. Inspection of the set of HGMD mutations in disordered regions revealed that a substantial fraction (438 out of the total of 1110) are in collagen. At first glance, it seems odd that collagen should be classed as a disordered protein, but this is a correct characterization - the bulk of collagen molecules are formed from a homo-triple helix. A hypothetical monomer would be structurally disordered, and the repeat triplet of the sequence is one of the possible signatures of disordered regions. Nevertheless, from the point of view of this analysis, the collagen mutations are atypical of disordered regions in other proteins, so I again retrained the SNPs3D profile method, omitting these mutations. The PDF in disordered regions, omitting the

collagen mutations, is now substantially lower (0.69 versus 0.84). Figure 3-6A shows that result on monogenic disease mutations together with those from Polyphen2 (Adzhubei et al., 2010) and CADD (Kircher et al., 2014), omitting the collagen mutations. The results from all three methods are similar and show consistently lower PDFs for disordered versus disordered regions (0.85 - 0.89 in ordered regions, 0.69 - 0.68 in disordered regions). That is, for monogenic disease, the fraction of mutations predicted deleterious is about 20% lower for disordered than ordered regions. The reason for this is unclear, but it is likely that the feature sets used in the sequence methods are not an optimal choice for disordered regions, and result in a higher false negative rate. If so, this will depress the values for cancer mutations in disordered regions as well.

Indeed, Figure 3-6B shows that the fraction of driver gene mutations predicted deleterious in disordered regions is consistently about 30% lower than in ordered regions. The difference here is larger than for monogenic disease (30% versus 20%), suggesting that even after allowing for the apparent high false negative rate in disordered regions, these may contain a lower fraction of driver mutations than ordered regions, consistent with the results in Figure 3-5.

Figure 3-6: (**A**) The fraction of predicted deleterious mutations is approximately 20% lower in disordered regions of monogenic disease proteins than in ordered regions. (**B**) For cancer driver proteins, the fraction of predicted deleterious mutations in disordered regions is approximately 30% lower than for ordered regions. (**C**) Fraction

of predicted deleterious mutations in the ordered (circles and solid lines) and disordered (square and dashed lines) regions of COSMIC Cancer Gene Census driver proteins as a function of mutation recurrence. For mutations with low recurrence, the fraction of predicted deleterious mutations is consistently lower in disordered regions than in ordered regions. Both fractions rise as a function of mutation recurrence and converge when mutations are observed for more than 10 times. Error bars show 95% confidence intervals derived from 100 rounds of bootstrapping.



(Figure 3-6, see above for caption.)

3.3.7 Other properties of mutations in disordered versus ordered regions

We found no tendency for tumor suppressors and oncogenes to be differently distributed between disordered and ordered regions. Similarly, we found no tendency for monogenic disease mutations in genes classified as dominant versus recessive to be differently distributed in disordered and ordered regions.

3.3.8 Protein surface and core mutations

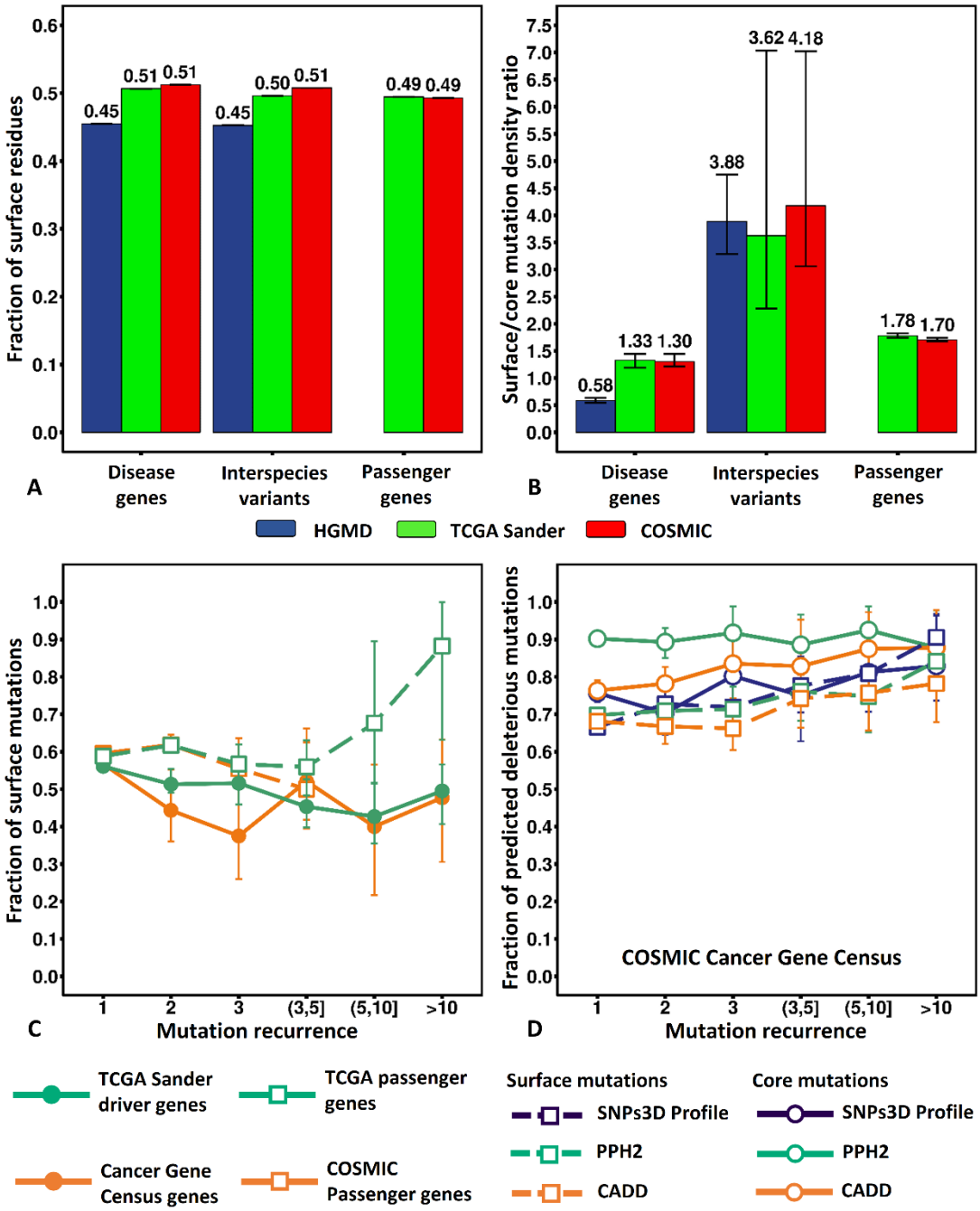
A second structural feature that provides insight into the mutation properties is the fraction of mutations on the surface of proteins versus in the core. Figure 3-7A shows that the fraction of all residues designated 'surface' (using STRIDE (Eisenhaber & Argos, 1993; Eisenhaber et al., 1995; Frishman & Argos, 1995), see Methods) is similar for all categories of protein, at approximately 50% (See Figure 3-8A for the mutations in the core). As shown in Figure 3-7B, the relative density of monogenic disease mutations on the surface to that in the core is only 0.58, showing a strong tendency for mutations in this class of disease to be buried (See Figure 3-8B for the mutations in the core). In contrast to this, both cancer driver gene sets show an enrichment of 1.3 for mutations on the surface versus in the core. Passenger gene mutations show a larger surface enrichment of 1.75 on average, and by far the highest surface enrichment is for inter-species variants, 3.6 to 4.2. The latter values reflect the fact that there are more possible neutral mutations on the surface than in the interior, so substitutions are more likely to be fixed on the surface, and there are more opportunities for benign passenger mutations there as well. We therefore expect some of the higher relative density on the surface of driver proteins arises from the

accumulation of passengers. However, unlike the data for disordered regions, the fraction of driver gene mutations on the surface as a function of mutation recurrence does not show a significant trend (Figure 3-7C) and so does not support an excess of surface passengers. Interpretation of this plot is complicated by a strong tendency for oncogene mutations to be on the surface and tumor suppressors to be in the core (see below). Since oncogene mutations have a higher recurrence rate than tumor suppressors, that tendency will dampen any relationship between surface and recurrence. (The corresponding surface density versus mutation load plot does show the expected relationship, but because of limited structural data, 95% confidence limits are large - data not shown). An alternative probe of the extent of passengers in surface regions is to consider the fraction of surface mutations predicted deleterious as a function of recurrence, since this fraction is similar for tumor suppressors and oncogenes (Table 3-5). Figure 3-7D shows a strong trend of increasing deleterious rate with mutation recurrence, consistent with the results from the passenger proteins and interspecies variant density results. Overall, the data support the conclusion that a substantial part of the excess surface mutations in cancer driver proteins are passenger mutations.

I also examined the surface to core distribution for mutations in oncogenes and tumor suppressors (Figure 3-9). Unlike the corresponding data for disorder/order, there is a marked difference in surface enrichment for the two classes of genes: for oncogenes the density of surface mutations is 1.9 times that of core mutations, while for tumor suppressors, the density is lower on the surface than in the core (average 0.85 that of

the core). For monogenic disease, there is also a smaller but still significant difference between the surface to core densities for genes classified as dominant and those classified as recessive, (density ratio of 0.68 for dominant versus 0.45 for recessive).

Figure 3-7: (**A**) For all classes of protein, about half of all residues are designated surface. (**B**) Ratio of mutation density on the surface to that in the core. The density of monogenic disease mutations on the surface is only about ½ that in the core, whereas for cancer driver protein mutations the density is higher on the surface. High ratios for interspecies variants and passenger proteins are consistent with less functional constraints on surface residues. (**C**) The fraction of surface mutations in cancer driver genes does not significantly correlate with mutation recurrence, so does not support an excess of surface mutations being passengers. A confounding factor is the tendency for oncogene mutations to be on the surface. (**D**) Fraction of predicted deleterious mutations in the core (circles and solid lines) and on the surface (square and dashed lines) for COSMIC Cancer Gene Census driver proteins as a function of mutation recurrence. The fraction of predicted deleterious mutations on the surface rises from around 0.7 for single occurrence mutations to 0.8~0.9 for those occurring more than 10 times. Error bars show 95% confidence intervals derived from 100 rounds of bootstrapping.



(Figure 3-7, See above for caption.)

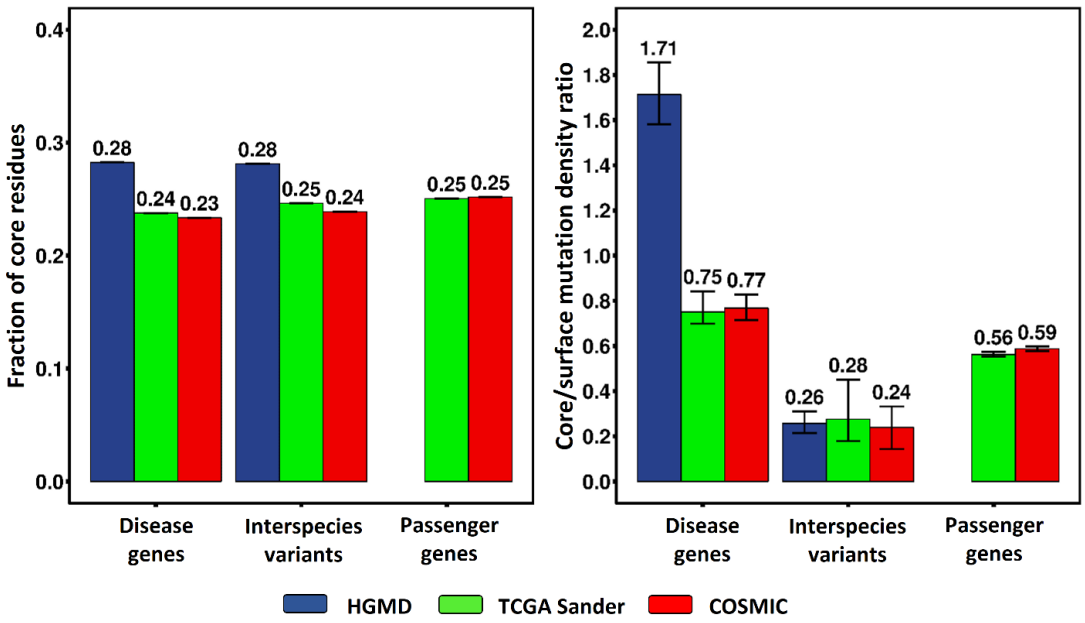


Figure 3-8 (A) Fractions of core residues and (B) ratio of mutation density in the core to that on the surface in monogenic disease genes and cancer driver genes, in the corresponding interspecies variants datasets, and in cancer passenger genes. The density ratio in HGMD disease genes is significantly higher than in the cancer driver genes. Compared to Figure 6, interspecies variants and somatic mutations in the passenger genes are more enriched on the protein surface, which supports that surface missense mutations are less deleterious and more tolerated.

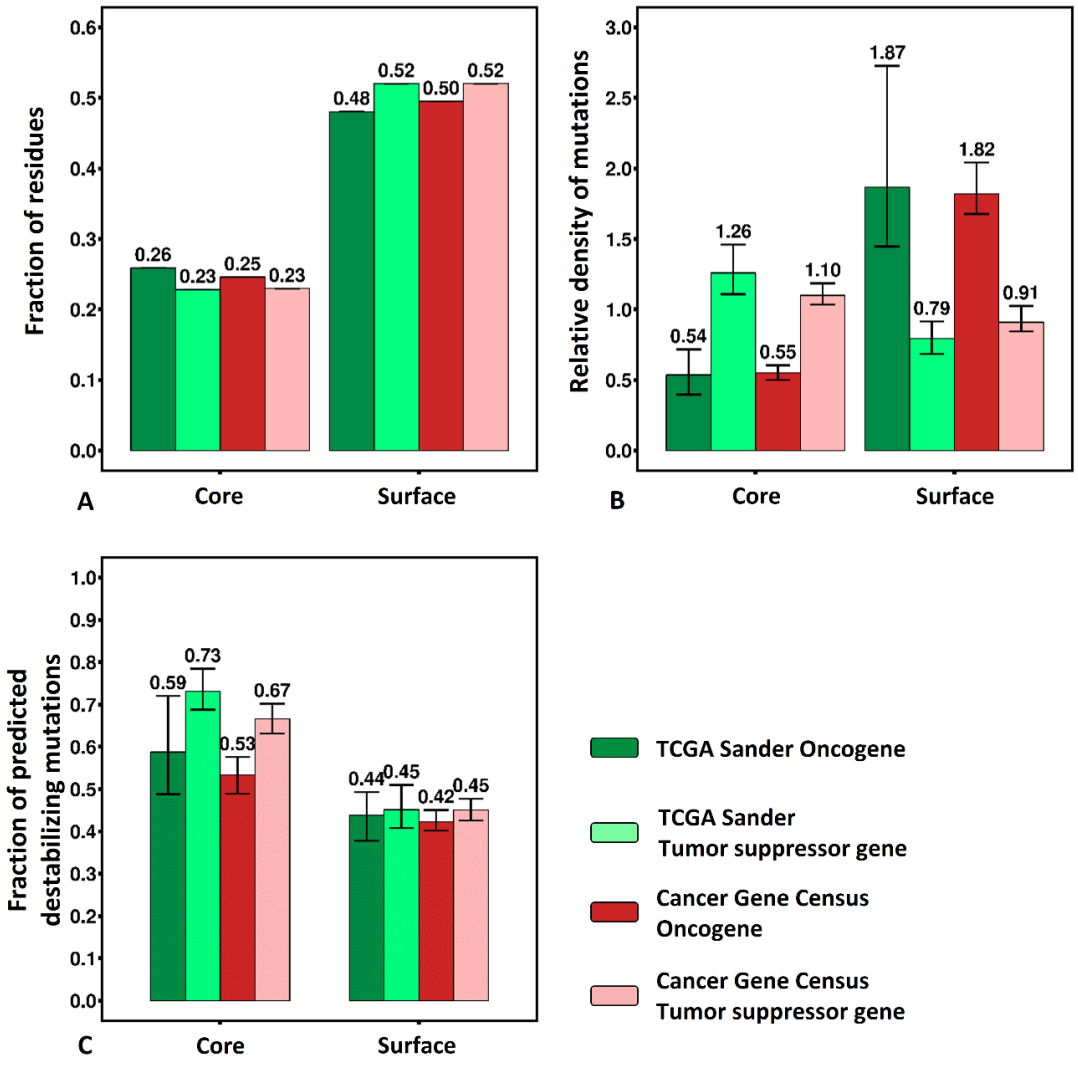


Figure 3-9 (**A**) The relative residue fraction and (**B**) the relative mutation density, and (**C**) the fraction of mutations predicted destabilizing using the SNPs3D Stability method for core and surface mutations in the TCGA Sander oncogene set, the TCGA Sander tumor suppressor gene set, the COSMIC Cancer Gene Census oncogene set, and the COSMIC Cancer Gene Census tumor suppressor gene set. Missense mutations are enriched in the core in tumor suppressor genes, and on the surface in oncogenes. In the core, the fraction predicted deleterious by SNPs3D Stability method is significantly higher in tumor suppressor genes than in oncogenes.

3.3.9 Role of structure destabilization

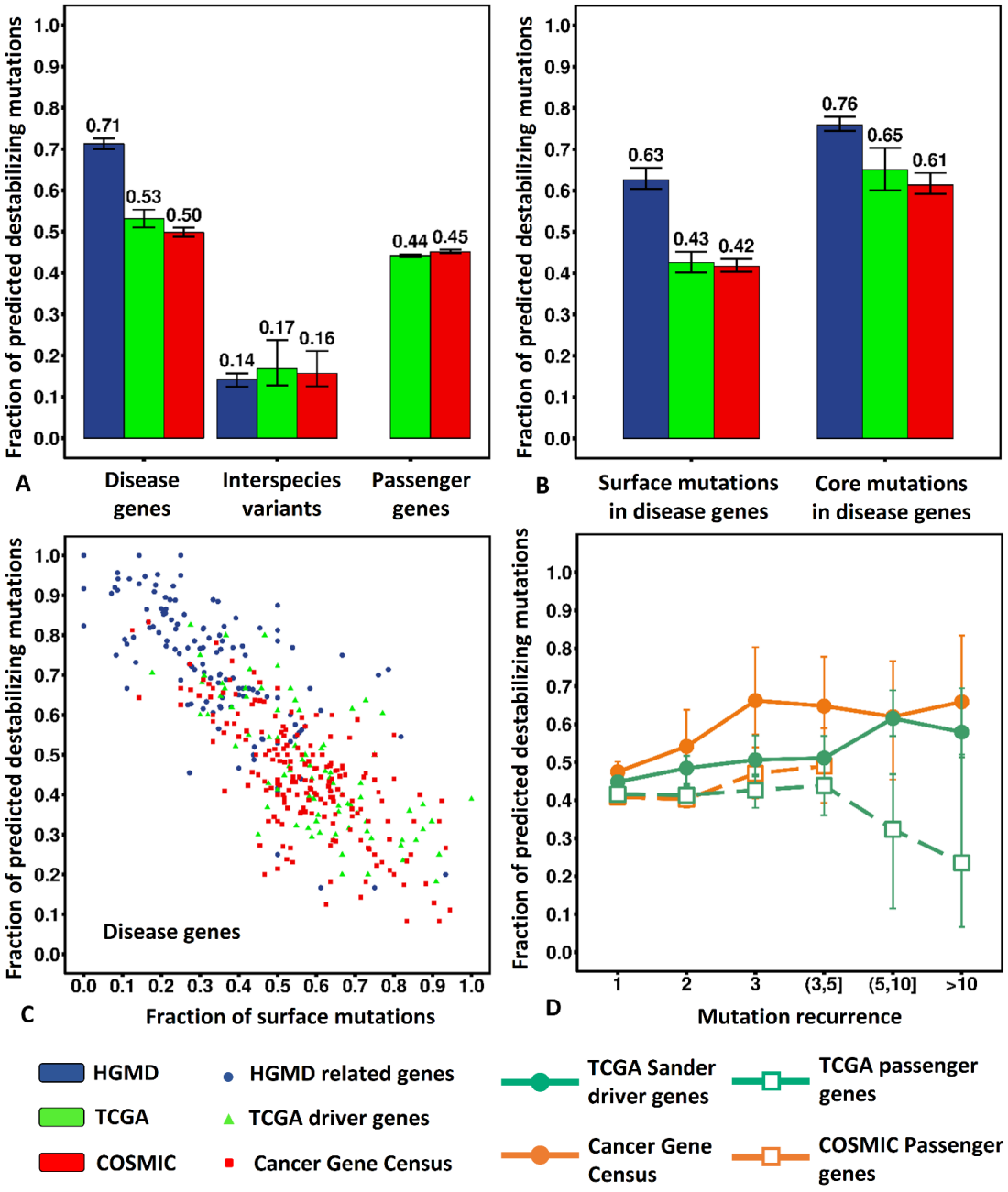
I next examine the role of structure destabilization in cancer mutations compared to those in monogenic disease. As noted earlier, destabilization plays a major role in monogenic disease mechanisms (Yue et al., 2005) and our earlier analysis suggests a significant role in cancer too (Shi & Moult, 2011). For this purpose, I trained separate stability SVMs on surface and core monogenic disease mutations, with interspecies variants in those regions as controls. Figure 3-10A shows that overall 0.71 of monogenic disease mutations are predicted to be destabilizing (similar to the value Yue & Moult reported earlier (Yue & Moult, 2006)), whereas only about 0.50 to 0.53 cancer driver gene mutations are predicted destabilizing, a large difference. The interspecies variant results provide an estimated false positive rate of 0.15. To address possible bias arising from training the stability method on monogenic disease data, I retrained using cancer data. This model was unsatisfactory in that it delivered much higher false positive rates (0.24 to 0.41 of interspecies variants predicted destabilizing, Table 3-5), but the relationship between the fraction of monogenic disease mutations predicted destabilizing and the fraction for cancer driver genes is similar to that obtained with the monogenic disease model (0.81-0.84 for monogenic disease and 0.64-0.68 for cancer driver gene mutations, Table 3-5), supporting the conclusion that destabilization rates are substantially higher for monogenic disease than for cancer driver gene mutations.

Based on the other analyses, we expect that passenger mutations in driver genes are one contributor to the lower fraction predicted destabilizing there. Examination of the

fraction of destabilizing mutations as a function of mutation recurrence (Figure 3-10D) supports a role for this factor, although less strongly than in the corresponding sequence analysis. A second contribution to different levels of destabilization in cancer and monogenic disease may come from the higher proportion of surface mutations in cancer: surface mutations are intrinsically less likely to be destabilizing. To help isolate this effect, I examined the role of destabilization in surface and core mutations separately. As expected, values for surface and core are markedly different: For monogenic disease, 0.76 of mutations in the core are predicted destabilizing, compared with 0.63 for surface. For cancer driver gene mutations, average values are 0.63 for core and 0.42 for surface (Figure 3-10B).

We expected that the distinction between surface and core destabilization properties might be particularly sensitive to whether a mutation is an oncogene or a tumor suppressor, so also examined the surface/core properties of these two classes separately. Indeed, tumor suppressor destabilizing fractions are higher in the core than those for oncogenes (0.73/0.67 versus 0.59/0.53), while surface values are similar for the two classes of mutation (Table 3-5, Figure 3-9). Thus, particularly for tumor suppressor mutations, a high (>70%) fraction of core mutations in both monogenic disease and cancer destabilize protein structure.

Figure 3-10: Fraction of predicted destabilizing mutations. (**A**) More than 70% of monogenic disease mutations are predicted destabilizing compared with only about half of mutations in cancer driver genes. The value for passenger gene mutations is not much lower than for driver genes. (**B**) The predicted destabilizing fraction is substantially lower on the surface than in the core for both types of disease. (**C**) Dependence on the fraction of mutations predicted destabilizing on the fraction of surface mutations. Each point is for one gene. The fewer mutations on the surface, the higher the fraction that predicts destabilizing, and monogenic disease genes (blue) tend to have a lower surface fraction than cancer genes. (**D**) Relationship between the fraction predicting deleterious and recurrence for cancer mutations. The driver gene destabilizing fraction increases with recurrence, consistent with a mixture of driver and passenger mutations in these genes. There is no trend for passenger gene mutations. Results are for SVMs trained on monogenic disease surface and core mutations separately. Error bars show 95% confidence intervals derived from 100 rounds of bootstrapping.



(Figure 3-10, See above for caption)

3.4 Discussion

In this paper, I have used computational methods together with sequence and structure information to investigate and compare the properties of missense mutations causative of monogenic disease and driving cancer. The principal findings are as follows:

3.4.1 Most monogenic disease and cancer driver mutations are under selection pressure, and so can be identified with sequence-based methods

After allowing for the effects of passenger mutations in cancer driver genes, I find a high fraction (> 80%) of mutations causing monogenic disease and of cancer driver mutations are predicted to be deleterious. These results are consistent across three different methods trained on monogenic disease (Figures 3-3, Figure 3-4). There are two primary implications. First, sequence methods trained on monogenic disease data are effective at identifying cancer drivers. Second, the large majority of mutations positions in both types of disease are under strong selection pressure (otherwise the sequence methods used would not be effective). While this was likely for most monogenic diseases, which are often severe and early onset, it is less obvious for cancer mutations, selected within a clone primarily to promote cell growth. It is not yet clear what fraction of the apparent false negatives - disease mutations not predicted deleterious - are technical false negatives or mutations that do not affect fitness.

3.4.2 Mutations in disordered regions play a limited role in both monogenic disease and cancer

Figure 3-5A shows that there is a much higher involvement of disordered regions in cancer than in monogenic disease, with only 10% of monogenic disease mutations in these regions, compared with 31-34% of cancer driver protein mutations. Two factors contribute to this difference. First, cancer driver genes are unusual in containing almost three times as much disorder as monogenic disease genes or passenger genes. That higher disordered fraction likely reflects a different functional spectrum for these proteins. In particular, it has been noted that these proteins are more hub-like (Goh et al., 2007; Jonsson & Bates, 2006; Kar, Gursoy, & Keskin, 2009) (involved in interactions with many partners), perhaps implying that more disordered regions are required to provide specificity for a range of protein binding partners (Fornili, Pandini, Lu, & Fraternali, 2013; J. Liu, Faeder, & Camacho, 2009). For example, the intrinsically disordered terminal trans-activation domain of P53 binds to three different protein partners in three different conformations (Oldfield et al., 2008). Second, the relative density of driver protein mutations in these regions is twice as high as for mutations in monogenic disease genes (~0.64 versus 0.30). However, the even higher relative mutation density in passenger protein disordered regions (~0.9) and for interspecies variants (~2.6) suggests that part of the cancer excess density is a consequence of benign passenger mutations being more likely to lie in disordered regions. Analysis of the relative densities as a function of the mutation recurrence and cancer mutational load confirm this is the case. The fraction of mutations predicted deleterious in disordered regions of cancer driver proteins is also low compared with

that for monogenic disease, also consistent with a high fraction of passengers in these regions (Figure 3-6). As noted above, the apparently deleterious mutations in disordered regions may often be involved in protein-protein interactions. The extent to which disordered regions of proteins are involved in function has not been clear (Vacic et al., 2012). Two aspects of these results confirm that disordered regions are much less functionally significant - the very low fraction of monogenic disease mutations there, and the high concentration of interspecies variants and passenger mutations.

3.4.3 Cancer oncogene mutations tend to be on the protein surface, whereas monogenic disease mutations and tumor suppressors tend to be in the core

The surface density of mutations in cancer driver genes is higher than in the core, and for mutations in oncogenes, it is nearly twice as high. While some of this difference reflects excess passenger mutations on the surface, it also likely reflects the greater role for disruption of intermolecular interactions in cancer (Nishi et al., 2013) and also that gain of function oncogene mutations tend to affect surface processes such as kinase conformational states related to phosphorylation (Blume-Jensen & Hunter, 2001). Conversely, tumor suppressor mutation density is higher in the core than on the surface, and for monogenic disease mutations, the core density is twice that of the surface. As discussed below, these values reflect the high role of structure destabilization for these classes of mutation. In monogenic disease, there is a higher relative density of surface mutations in autosomal dominant genes than recessive ones (0.68 versus 0.45). A number of mechanisms are involved in autosomal dominant

disease, including haplo-insufficiency, oligomer structure (of which collagen mutations are an example (Lamandé et al., 1998)), and gain of function. The latter mechanism likely contributes most to the surface/core signal, in a manner analogous to that of gain of function mutations in oncogenes. Examples are of monogenic surface gain of function mutations in the calcium sensing receptor (CASR), causing hypocalcemia, and in Luteinizing Hormone/Choriogonadotropin Receptor (LHCGR), causing Familial Male-Limited Precocious Puberty (FMPP).

3.4.4 A large fraction of monogenic disease and cancer tumor suppressor mutations in the protein core destabilize protein structure

Approximately ¾ of both monogenic disease mutations and cancer tumor suppressor mutations in the protein core are predicted to destabilize protein structure. A prediction of destabilization using this method is equivalent to a major decrease in protein abundance *in vivo* (Yue et al., 2005), either through misfolding or reduced protein half-life. Most monogenic disease missense mutations result in a major loss of molecular function (for example, (Shi, Sellers, & Moult, 2012) and (Yin, Kundu, Pal, & Moult, 2017)). To the extent that core tumor suppressor mutations can be considered to represent all driver mutations, the result implies that in this class of disease too there is usually major loss of protein function, rather than a subtle effect at that level.

What about the other 25% of core monogenic mutations and tumor suppressor mutations which do not appear to destabilize structure? Are these more subtle in their

effect on protein function? Recessive mutations provide some evidence here, since a disease outcome for most of these involves a 50% or greater loss of molecular function. The predicted destabilization fraction for the recessive mutations is approximately the same as for monogenic disease as a whole. That suggests that most of the remaining 25% are a combination of false negatives of the computational method and mechanisms other than destabilization, rather than mutations with a subtle effect on function.

There are some oncogene mutations in the core region, and about 50% of these are predicted to destabilize protein structure, at first glance a surprising result, since these should be gain of molecular function. As noted earlier, less than 1/3 of oncogene mutations are in the core, so that a 50% destabilization rate corresponds to just 1/6 of oncogene mutations. The estimated false positive rate is 0.15, close to that value, and there may be some cases where oncogene gain of function is the result of destabilization of a regulatory domain. Also, the definition of oncogenes and tumor suppressors is not always unambiguous. In compiling the oncogene and tumor suppressor lists I noted that 10 genes had been classified as oncogenes by one group and tumor suppressors by the other (these were excluded). There are also examples where a gene may behave as an oncogene in some circumstances and a tumor suppressor in others (Manfredi, 2010).

3.4.5 Mutations in passenger genes show a high fraction of deleteriousness

A surprisingly high fraction (more than 50%) of mutations in passenger genes appear deleterious with the sequence methods used here. The estimated false positive rate is much lower (10% or less). Stability analysis supports this observation, with almost as high a fraction of passenger gene mutations predicted destabilizing as in driver genes. There are at least two possible explanations. One is that there is insufficient selection pressure to eliminate these mutations in a typical cancer. Simulations of cancer progression suggest that moderately deleterious mutations will escape elimination by various population genetics mechanisms, and so accumulate, sometimes impending cancer progression (McFarland et al., 2013). The other is that there is a significant concentration of unrecognized driver genes. As noted earlier, there is considerable variation in driver set definitions, so that it is expected this would be the case to some degree. But depending on the cancer type and particular case (Martincorena & Campbell, 2015), there may be up to 100s or even a thousand deleterious mutations spread across non-driver genes and a deleterious fraction of 0.5 implies that a substantial fraction of these mutations genes must be drivers or deleterious mutations not yet selected out, which seems improbable. The exact nature and impact of these mutations will repay further study.

In common with all analyses so far we have assumed a binary model of cancer drivers - a mutation is either a driver or not. But it may be that there is a continuous scale of driver impact with a few strong drivers and a long tail of mutations making secondary

contributions, loosely analogous to the contributions of variants to complex trait disease (Pritchard, 2001).

Chapter 4: Increasing the Stability of the Bacteriophage Endolysin PlyC Using Rationale-Based FoldX Computational Modeling

Published:

Heselpoth RD, Yin Y, Moult J., Nelson DC. 2015. Increasing the Stability of the Bacteriophage Endolysin PlyC Using Rationale-Based FoldX Computational Modeling. *Protein Engineering, Design & Selection* 28(4):85-92.

My contribution: computational experiments and data analysis

4.1 Abstract

Endolysins are bacteriophage-derived peptidoglycan hydrolases that represent an emerging class of proteinaceous therapeutics. While the streptococcal endolysin PlyC has been validated *in vitro* and *in vivo* for its therapeutic efficacy, the inherent thermosusceptible structure of the enzyme correlates to transient long-term stability, thereby hindering the feasibility of developing the enzyme as an antimicrobial. Here we thermostabilized the CHAP domain of the PlyCA catalytic subunit of PlyC using a FoldX-driven computational protein engineering approach. Using a combination of FoldX and Rosetta algorithms, as well as visual inspection, a final list of PlyC point mutant candidates with predicted stabilizing $\Delta\Delta G$ values was assembled and thermally characterized. Five of the eight point mutations were found experimentally to be destabilizing, a result most likely attributable to computationally modeling a

complex and dynamic nine-subunit holoenzyme with a corresponding 3.3-Å X-ray crystal structure. However, one of the mutants, PlyC (PlyCA) T406R, was shown experimentally to increase the thermal denaturation temperature by ~2.2 ℃ and kinetic stability 16 fold over wild-type. This mutation is expected to introduce a thermally advantageous hydrogen bond between the Q106 side-chain of the N-terminal GyH domain and the R406 side-chain of the C-terminal CHAP domain.

4.2 Introduction

Endolysins, also termed phage lysins or enzybiotics, are bacteriophage-encoded peptidoglycan hydrolases (Nelson et al., 2006). During a lytic bacteriophage (phage) replication cycle within the host bacterium, endolysins are expressed and accumulate in the cytosol in a fully folded and active conformation. The exact moment of cell lysis is then highly regulated by holins, hydrophobic membrane proteins that generate pore-forming complexes on the cytoplasmic membrane, providing cytosolic endolysins access to their peptidoglycan substrate (I. N. Wang, Smith, & Young, 2000; Young, 1992). The endolysin then degrades the peptidoglycan upon direct contact due to the hydrolysis of key covalent bonds within the cell wall structure, resulting in osmotic lysis and liberation of intracellular progeny virions. With this mechanistic understanding, the exogenous application of a purified recombinant endolysin to susceptible Gram-positive bacteria produces the same bacteriolytic phenotype without the presence of the bacteriophage or holins and thus represents an alternative antimicrobial to treat antibiotic-resistant bacterial infections (Fischetti, Nelson, & Schuch, 2006).

PlyC is an endolysin derived from the streptococcal C_1 lytic phage that has been validated *in vitro* for its bacteriolytic efficacy against groups A (GAS), C (GCS) and E (GES) streptococci and *in vivo* for its ability to protect mice from streptococcal challenge (Krause, 1957; Nelson et al., 2001). When added to GAS (*Streptococcus pyogenes*) *in vitro*, 10 ng of PlyC was able to cause a 7 log decrease in colony forming units in 5 s, making this endolysin ~100 fold more active than any other characterized endolysin to date (Nelson et al., 2001). Unlike other endolysins, which are single gene products consisting of one or more enzymatically active domains (EAD) and a cell wall binding domain (CBD), PlyC consists of a novel multimeric structure with nine distinct subunits (McGowan et al., 2012; Nelson et al., 2006). Eight identical PlyCB monomers interact to form a symmetrical octameric ring structure that serves as the CBD of the holoenzyme. The ninth subunit, PlyCA, functions as the EAD of the endolysin and consists of three domains. The catalytically-active N-terminal glycosyl hydrolase (GyH) and C-terminal cysteine, histidine-dependent amidohydrolase/peptidase (CHAP) domains act together synergistically to generate the robust bacteriolytic mechanism of the enzyme, whereas the central helical docking domain interacts with the PlyCB CBD to promote the formation of the holoenzyme structure (McGowan et al., 2012).

Thermal denaturation of PlyC by means of differential scanning calorimetry (DSC) showed that the PlyCB octamer is endogenously thermostable, displaying a thermal transition temperature (T_G) of 75.0 ℃, whereas the PlyCA EAD is thermosusceptible, with a T_G of 46.2 ℃ (F. Schwarz, personal communication). While the dissociation of

the PlyCB octamer into isolated monomers is a reversible thermodynamic process, the unfolding of the individual PlyCA and PlyCB monomers is an irreversible event, which is supported by their inability to refold after being heat-denatured. The C-terminal CHAP domain of PlyCA was shown to have a T_G of 39.1 ℃ when isolated, compared to a T_G of 46.0°C associated with PlyCAΔCHAP in a PlyCΔCHAP background (i.e. PlyC holoenzyme with a PlyCA C-terminal CHAP domain deletion), suggesting that the CHAP domain of PlyCA is the most heat-labile structural component of the PlyC holoenzyme.

Although the number of thermodynamically characterized endolysins is limited, there are examples of endolysins that display similar structural instability to that of PlyC. For example, the *Staphylococcus aureus* endolysin LysK as well as the *Streptococcus pneumoniae* endolysins Cpl-1, Pal and Cpl-7 are devoid of activity or unfold at 42.5 ℃, 43.5 ℃, 50.2 ℃ and 50.4 ℃, respectively (Bustamante, Rico-Lastres, Garcia, Garcia, & Menendez, 2012; Filatova, Becker, Donovan, Gladilin, & Klyachko, 2010; Sanz, Garcia, Laynez, Usobiaga, & Menendez, 1993; Varea et al., 2004). In congruence to the Arrhenius equation, the thermolability of PlyC and other endolysins correlates to a short-term therapeutic shelf-life expectancy (Anderson & Scott, 1991).

A number of computational methods have proven partially effective at identifying single amino acid substitutions that result in increased thermodynamic stability of a protein (Cheng, Randall, & Baldi, 2006; Gilis & Rooman, 2000; Guerois et al., 2002;

Parthiban, Gromiha, Hoppe, & Schomburg, 2007; Parthiban, Gromiha, & Schomburg, 2006; Schymkowitz, Rousseau, et al., 2005; Zhou & Zhou, 2002). One example is FoldX (Guerois et al., 2002; Schymkowitz, Borg, et al., 2005; Schymkowitz, Rousseau, et al., 2005), which uses an empirical potential derived from a weighted combination of physical energy terms (e.g. van der Waals interactions, hydrogen bonding, electrostatics and solvation), statistical energy terms and structural descriptors. In third-party testing, FoldX has been shown to perform with useful accuracy across all protein structure types, yielding a correlation coefficient of 0.5 between estimated and experimental $\Delta\Delta G$ (Khan & Vihinen, 2010; Potapov, Cohen, & Schreiber, 2009). Rosetta was developed primarily for designing proteins with desirable properties, including new protein folds (Brian Kuhlman et al., 2003), novel enzymatic activity (Jiang et al., 2008; Röthlisberger et al., 2008) and modified substrate specificity (Ashworth et al., 2006). The ddG module of Rosetta also provides a means of estimating $\Delta\Delta G$ for point mutations (Kellogg et al., 2011).

Here we aim to engineer enhanced stability of a thermolabile bacteriolytic enzyme using computational modeling. Using the PlyC holoenzyme structure as the template, our engineering strategy was to apply the FoldX and Rosetta algorithms together, in addition to subsequent visual inspection, to the C-terminal CHAP domain of PlyCA. By doing so, we were able to identify one point mutant, PlyC (PlyCA) T406R, which was shown experimentally to thermostabilize the PlyC holoenzyme structure and thereby enhance its long-term stability and therapeutic potential.

4.3 Materials and methods

4.3.1 Computational Modeling of PlyC Mutants

Initial atomic coordinates were taken from the PlyC holoenzyme X-ray crystal structure (Protein Data Bank ID 4F88). Due to the relatively low 3.3-Å resolution of the structure, polypeptide backbones and side-chains were adjusted using Rosetta Relax (Raman et al., 2009), followed by another round of side-chain orientation optimization using the FoldX3.0 RepairPDB command (Guerois et al., 2002; Schymkowitz, Borg, et al., 2005). The resulting coordinates were then processed with FoldX3.0 PositionScan to obtain estimated changes in folding free energy ($\Delta\Delta G_{FoldX}$) for all of the 2,945 possible CHAP domain point mutants (155 total CHAP domain residues multiplied by the 19 alternative natural amino acids). The structural environments of those mutations with a predicted $\Delta\Delta G_{FoldX} \leq$ -1 kcal/mol were then manually inspected to remove those judged likely to introduce unfavorable structural alterations. Finally, the remaining mutants were processed through the Rosetta ddg_monomer application (Kellogg, Leaver-Fay, & Baker, 2011) to yield the PlyC candidate mutant list for experimental study.

4.3.2 Bacterial Strains and Culture Conditions

S. pyogenes D471 (group A streptococcus) was maintained and grown in Todd Hewitt broth supplemented with 1% yeast extract as previously described (Nelson et al., 2001; Nelson, Schuch, Zhu, Tscherne, & Fischetti, 2003). *E. coli* strains DH5α and BL21(DE3)pLysS (Novagen) were grown in Luria-Bertani (LB) broth at 37 ℃ in a

shaking incubator unless otherwise stated. When needed, ampicillin (100 μg/ml) was added to the media.

4.3.3 Cloning and Site-directed Mutagenesis

The *plyC* operon was cloned into pBAD24 as previously described (Nelson et al., 2006). Site-directed mutagenesis was performed using the Phusion Site-Directed Mutagenesis Kit (Thermo Scientific). Mutations were introduced into the middle of the 30 nucleotide forward phosphorylated oligonucleotide primer for each mutant, with the reverse primer being complementary to the next 30 nucleotides upstream (Eurofins Scientific). The standard 50 μl PCR reaction mixture consisted of 1 ng of pBAD24::*plyC*, 1x Phusion HF Buffer, 0.2 mM dNTP, 0.5 μM of each primer and 1 U of Phusion DNA polymerase. The thermocycler heating conditions consisted of 98 ℃ for 30 s, 25x (98 ℃ for 10 s; 65 ℃ for 30 s; 72 ℃ for 4 min) and 72 ℃ for 5 min. The resulting PCR products were then ligated and transformed into *E. coli* DH5α. Plasmid DNA was extracted from successful transformants and mutations were confirmed by nucleotide sequencing (Macrogen USA). Vector constructs comprising an insert with the correct nucleotide sequence were transformed into *E. coli* BL21(DE3)pLysS for protein expression.

4.3.4 Protein Expression and Purification

E. coli BL21(DE3)pLysS harboring the wild-type and mutant pBAD24::*plyC* expression constructs were grown to mid-log phase in 1.5L LB supplemented with

ampicillin in a 4L baffled Erlenmeyer flask. Protein expression was induced with 0.25% L-arabinose at 37 ℃ overnight. The cells were harvested the following morning at 7,000 RPM, resuspended in phosphate buffered saline (PBS), pH 7.25, supplemented with 1 mM phenylmethanesulfonyl fluoride (Sigma-Aldrich) and sonicated on ice for 15 min. The insoluble cell debris from the cell lysate was pelleted at 13,000 RPM for 1 h at 4 ℃. The soluble endolysins were then purified as previously described (Nelson et al., 2001). Protein solubility and purity were assessed on a 4-15% gradient sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) gel (Laemmli, 1970).

4.3.5 *In Vitro* Endolysin Activity on *S. pyogenes*

Spectrophotometric-based turbidity reduction assays were performed to determine the bacteriolytic activity of each endolysin investigated. An overnight culture of *S. pyogenes* D471 was harvested at 4,000 RPM for 15 min, washed once with PBS, pH 7.2, and resuspended to an OD_{600} = 2.0. In a flat-bottomed 96-well plate, the purified endolysin at an initial concentration of 8.84 μM (1 mg/ml) was serial diluted in 100 μl of PBS buffer. An equal volume of bacteria was then mixed with the different enzyme concentrations and the OD_{600} was monitored kinetically on a SpectraMax 190 microplate spectrophotometer (Molecular Devices) every 6 s for 30 min at 37 ℃. The amount of time (s) to decrease the initial OD_{600} by 50% was then plotted against the enzyme molar concentration and fit with a one-phase exponential decay curve. 1 U of enzyme activity was equated to the amount of endolysin (μg) required to decrease the

OD_{max} by 50% in 15 min. Each independent turbidity reduction assay was performed in triplicate.

4.3.6 Circular Dichroism Spectroscopy

A Chirascan CD spectrometer (Applied Photophysics) equipped with a thermoelectrically controlled cell holder was used for all CD experiments. For secondary structure far-ultraviolet (UV) analysis, the endolysins were at a 0.1 mg/ml concentration in 20 mM sodium phosphate buffer, pH 7.0. CD spectra were obtained in the far-UV range (190-260 nm) in a 1 mm path length quartz cuvette at 1 nm steps with 5-second signal averaging per data point. Spectra were collected in triplicate, followed by averaging, baseline subtraction, smoothing and conversion to mean residue ellipticity (MRE) by the Pro-Data software (Applied Photophysics). Secondary structure prediction was performed using the Provencher and Glockner method provided by DICHROWEB (Provencher & Glockner, 1981; Whitmore & Wallace, 2004). Melting experiments were performed by heating the endolysins at a 1 mg/ml concentration in 20 mM sodium phosphate buffer, pH 7.0, from 20 ℃ to 95 ℃ at 1 ℃/min. MRE was monitored at 222 nm in a 1 mm path length quartz cuvette at 0.5 ℃ steps with 5-second signal averaging per data point. The melting data was smoothed, normalized and fit with a Boltzmann sigmoidal curve. The first derivative of the melting curve was then taken to determine the temperature, T_G, at which the concentration of the folded and unfolded states of the PlyCA subunit were the same. This temperature was defined as the minimum in the first derivative graph (Fallas & Hartgerink, 2012).

4.3.7 Differential Scanning Calorimetry

DSC experiments were performed on a Nano DSC differential scanning calorimeter (TA Instruments) at a constant pressure of 3 atm. All samples were degassed for at least 15 minutes prior to the experiment. The sample and reference cells consist of an optimal operational volume of 0.3 ml and were calibrated with equal volumes of 20 mM sodium phosphate buffer, pH 7.0, by means of three consecutive heating/cooling cycles from 15 ℃ to 105 ℃ and 105 ℃ to 15 ℃ at 1 ℃/min. The endolysins were then heated from 15 ℃ to 105 ℃ at a 1 ℃/min heating rate in 20 mM sodium phosphate buffer, pH 7.0, using a final protein concentration of 1 mg/ml followed by immediate cooling from 105 °to 15 ℃ at 1 ℃/min. Data analysis by means of baseline subtraction and curve fitting was performed by the NanoAnalyze software (TA Instruments). Due to a scan rate-independence displayed by PlyC during calorimetric analysis (F. Schwarz, personal communication), equilibrium thermodynamics were applied to the finalized calorimetric dataset. The thermal transition temperature, T_G, was defined as the mid-point of each thermal transition.

4.3.8 45 ℃ Kinetic Stability Assay

The various endolysins investigated were incubated in a 45 ℃ hot plate in PBS, pH 7.2, at a 44 nM (5 μg/ml) concentration for a total of 3 hours. At 20 minute increments, a 400 μl aliquot of the heated enzyme was removed and incubated on ice for 5 minutes. Three adjacent wells of a 96-well plate were then filled with 100 μl of the cooled enzyme, followed by the addition of an equal volume of *S. pyogenes* D471 (see *In Vitro* Endolysin Activity on *S. pyogenes* for cell preparation). The residual

lytic activity of the endolysin was analyzed via turbidity reduction assay by monitoring the OD_{600} every 6 s for 20 min. The activity of the endolysin was equated to the V_{max} (milli-OD units per min) corresponding to the linear portion of the resulting killing curve. Residual lytic activity was normalized to the activity displayed in the absence of heat treatment.

4.4 Results

FoldX was applied to the C-terminal CHAP domain of PlyCA (CHAP is comprised of PlyCA amino acids 309-465; however, atomic coordinates were only available for residues 310-464), substituting each of the possible 19 alternative natural amino acids at each residue position, so generating a library of 2,945 PlyC mutants. Most of the mutations analyzed (n = 2,453) were predicted by FoldX to have either destabilizing or neutral effects on stability, resulting in a $\Delta\Delta G_{FoldX} \geq 0$ kcal/mol ($\Delta\Delta G_{FoldX} = \Delta G_{mut} - \Delta G_{wt}$) (Figure 4-1). All of the mutants (n = 92) that had a $\Delta\Delta G_{FoldX} \leq$ -1 kcal/mol were visually inspected, resulting in the elimination of another 61 mutants that appear to modify the PlyC structure in an unfavorable manner. Examples of these disadvantageous structure changes are disruption of salt-bridge and dipole interactions, replacement of salt-bridge interactions with weaker dipole interactions, generation of cavities in the hydrophobic core by the introduction of an amino acid with a smaller side-chain, exposure of hydrophobic side-chains at the surface, and disruption of the active site.

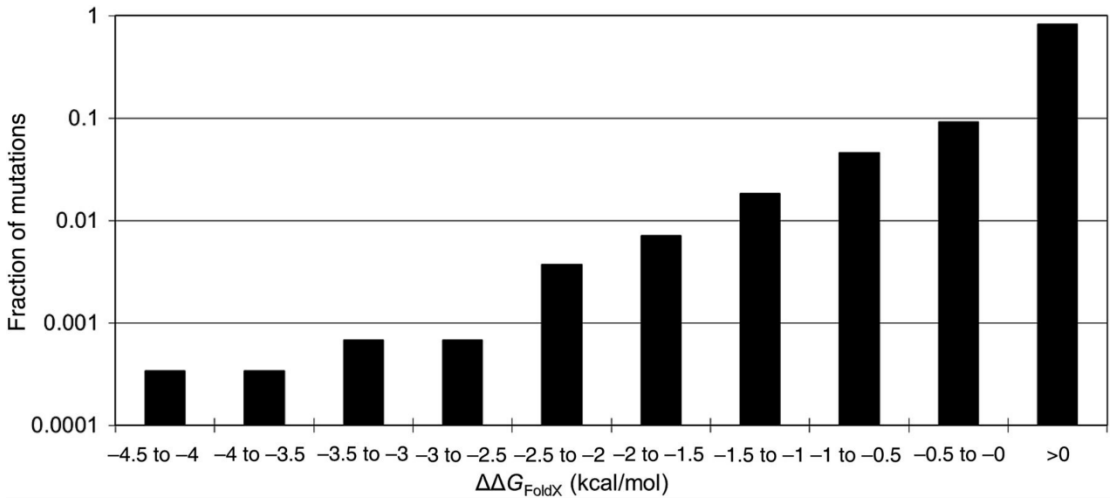


Figure 4-1. Log distribution of the predicted change in folding free energy ($\Delta\Delta G_{FoldX}$) for all 2,945 possible PlyCA CHAP domain point mutants calculated with FoldX 3.0 PositionScan. Mutations with $\Delta\Delta G_{FoldX} < 0$ are expected to increase protein stability. Only a small portion of mutations are predicted to be stabilizing.

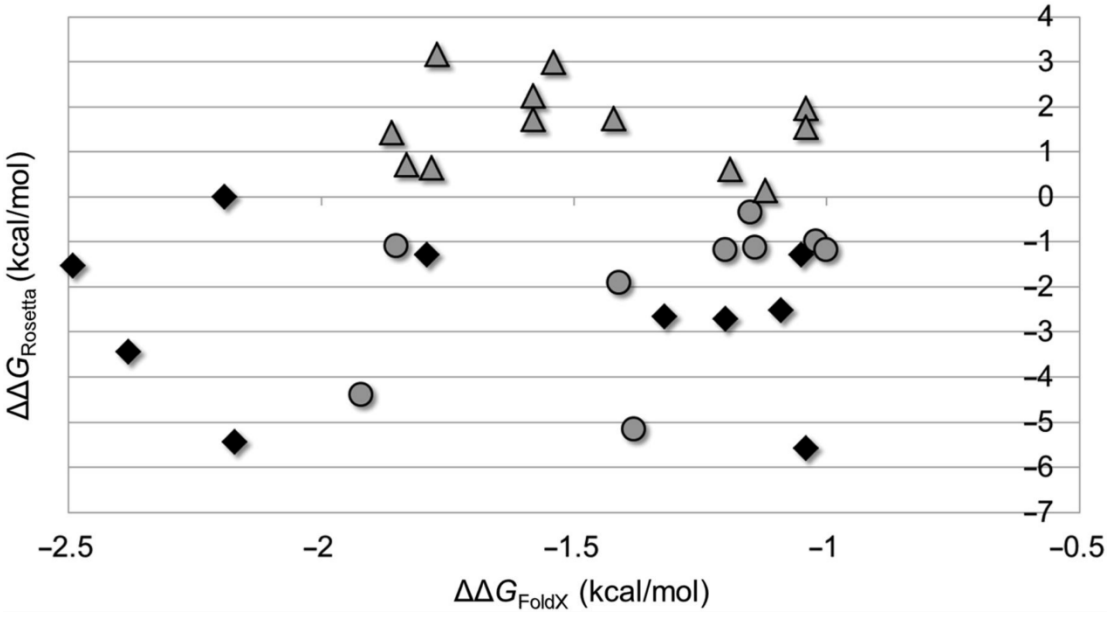
Table 4-1. List of the final 10 PlyC mutant candidates, including the specific mutation (column 1), the location of the mutation within the CHAP domain of PlyCA (column 2) and the calculated $\Delta\Delta G$ (kcal/mol) values by the FoldX (column 3) or Rosetta (column 4) algorithms.

PlyC Construct	Location	$\Delta\Delta G_{FoldX}$ (kcal/mol)	$\Delta\Delta G_{Rosetta}$ (kcal/mol)
Wild-type	--------	--------	--------
D330Y	Surface, near the active site	-1.09	-2.51
Q332H	Surface, near the active site	-2.19	0.01
Q332V	Surface, near the active site	-1.79	-1.29
C345T	Surface with potential domain-domain interaction	-1.20	-2.70
D375Y	Surface with potential domain-domain interaction	-2.49	-1.53
T381Y	Surface with potential domain-domain interaction	-1.32	-2.65
V384Y	Surface with potential intra-domain hydrogen bond	-1.04	-5.57
C404I	Hydrophobic core	-2.17	-5.44
T406R	Surface with potential domain-domain interaction	-1.05	-1.29
T421I	Hydrophobic core, near the active site	-2.38	-3.43

The impact of the mutations encoded by the remaining 31 mutants was analyzed by the Rosetta ddg monomer algorithm (Kellogg et al., 2011) to further evaluate likely stabilizing potential (Figure 4-2). Rosetta predicted 12 of the 31 mutants to be destabilizing ($\Delta\Delta G_{Rosetta}$ > 0.01 kcal/mol) and these were eliminated from further consideration (Figure 4-2, triangles). An additional mutant, PlyC (PlyCA) Q332H, has a predicted mildly destabilizing $\Delta\Delta G_{Rosetta}$ of 0.01 kcal/mol, but was retained since it is in an interesting location, adjacent to the active site cysteine of the CHAP domain, C333. Although mutations near the active site generally induce activity defects, there are several documented instances where such mutations improve

overall thermal stability (Daude, Topham, Remaud-Simeon, & Andre, 2013; Kamal, Mohammad, Krishnamoorthy, & Rao, 2012; Kanaya, Oobatake, & Liu, 1996; Lam, Yeung, Yu, Sze, & Wong, 2011; Xie et al., 2014; Zhi, Srere, & Evans, 1991). In addition to Q332H, the PlyC (PlyCA) mutants D330Y, Q332V, C345T, D375Y, T381Y, V384Y, C404I, T406R and T421I were also selected as candidates for experimental study, on the basis of predicted $\Delta\Delta G < 0$ kcal/mol by both of FoldX and Rosetta (Figure 4-2, diamonds, Table 4-1). These final 10 candidates consisted of mutations located near the CHAP domain active site (D330Y, Q332H, Q332V), in the hydrophobic core (C404I and T421I), and at the surface predicted to form an intra-domain hydrogen bond (V384Y) or an inter-domain interaction with the N-terminal GyH domain (C345T, D375Y, T381Y, T406R). The other nine mutants with FoldX and Rosetta $\Delta\Delta G < 0$ kcal/mol values were omitted from further characterization (Figure 4-2, circles). These had similar structural locations to those selected, and were hypothesized to employ analogous stabilizing mechanisms so that inclusion would not improve the diversity of the candidate pool.

Figure 4-2. Comparison between the $\Delta\Delta G_{FoldX}$ and $\Delta\Delta G_{Rosetta}$ values of the final 31 mutant candidates retained after manual curation. Twelve of these remaining mutants displayed a $\Delta\Delta G_{Rosetta} > 0.01$ kcal/mol and were not further considered (triangles). Of the remaining 19 mutants (circles and diamonds), 10 were selected for experimental characterization (diamonds). All of the final candidates had predicted $\Delta\Delta G \leq 0.01$ kcal/mol values by both FoldX and Rosetta algorithms.



(Figure 4-2, See above for caption)

4.4.1 Protein Solubility, Purity and Secondary Structure Determination

PlyC (PlyCA) mutants Q332H, Q332V, C345T, D375Y, V384Y, C404I and T406R all expressed as soluble holoenzymes and were purified to homogeneity based on SDS-PAGE analysis (Figure 4-3). No protein expression was observed for PlyC (PlyCA) D330Y and T381Y and therefore both were excluded from further characterization (data not shown). SDS-PAGE and far-UV CD secondary structure analysis of purified PlyC (PlyCA) T421I showed a mixed population of holoenzyme and uncomplexed PlyCB octamer structures (data not shown). To overcome this issue, a C-terminal 6x His-tag was added to PlyCA T421I. This mutant was expressed and purified in the same manner as the other PlyC mutants, with two alterations; protein expression was induced at 18 ℃ instead of 37 ℃, and there was an addition of

a final immobilized metal affinity chromatography (IMAC) step using a 5 ml Bio-Scale Mini Profinity IMAC Cartridge (Bio-Rad) to remove uncomplexed PlyCB octamers.

Protein secondary structure analysis was performed using far-UV CD. The CD spectra for all of the proteins analyzed were represented in terms of mean residue ellipticity (MRE) as a function of wavelength (Figure 4-4a). All eight of the purified PlyC mutants displayed no deviation in secondary structure when compared to that of wild-type. The far-UV spectra resembles that of an α/β folded protein, displaying ellipticity minima at 208 nm and 220 nm, and ellipticity maxima at 195 nm (Greenfield & Fasman, 1969; Kelly, Jess, & Price, 2005). Secondary structure composition analysis results depict highly homologous regular α-helical (±1.1%), distorted α-helical (±1.2%), regular β-strand (±1.0%), distorted β-strand (±0.5%), turn (±0.7%) and unordered (±0.7%) structures when comparing wild-type to the eight point mutants (data not shown). The normalized root mean square deviation (NRMSD) value, which measures the goodness-of-fit between back-calculated spectra (spectra extrapolated using the CONTIN method for soluble proteins with known crystal structures) and experimental spectra, for each sample was < 0.1, suggesting the back-calculated and experimental spectra are in close agreement. Thus, none of the point mutations introduced significantly affected the secondary structure of the holoenzyme.

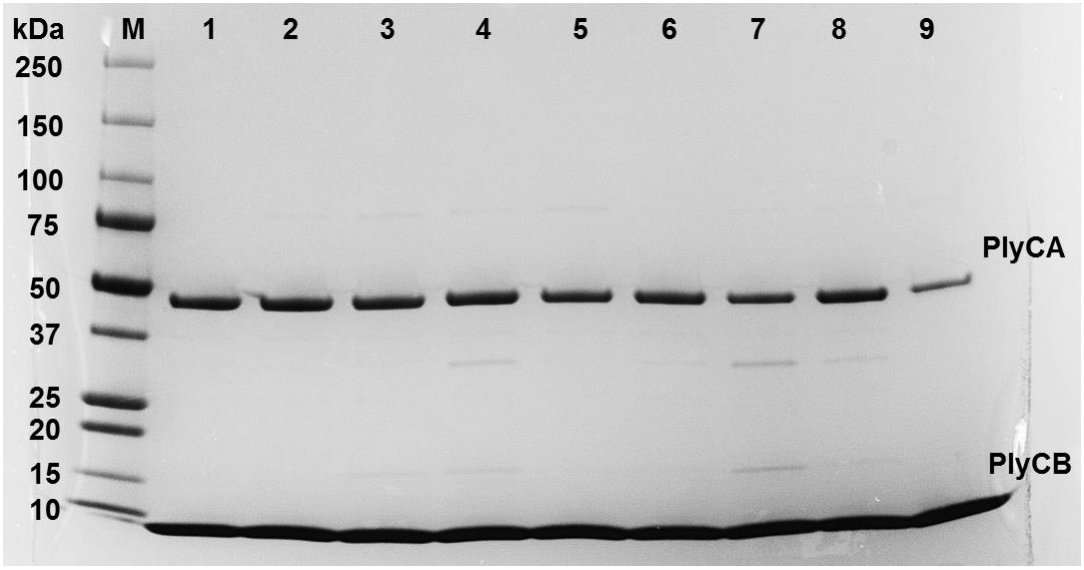
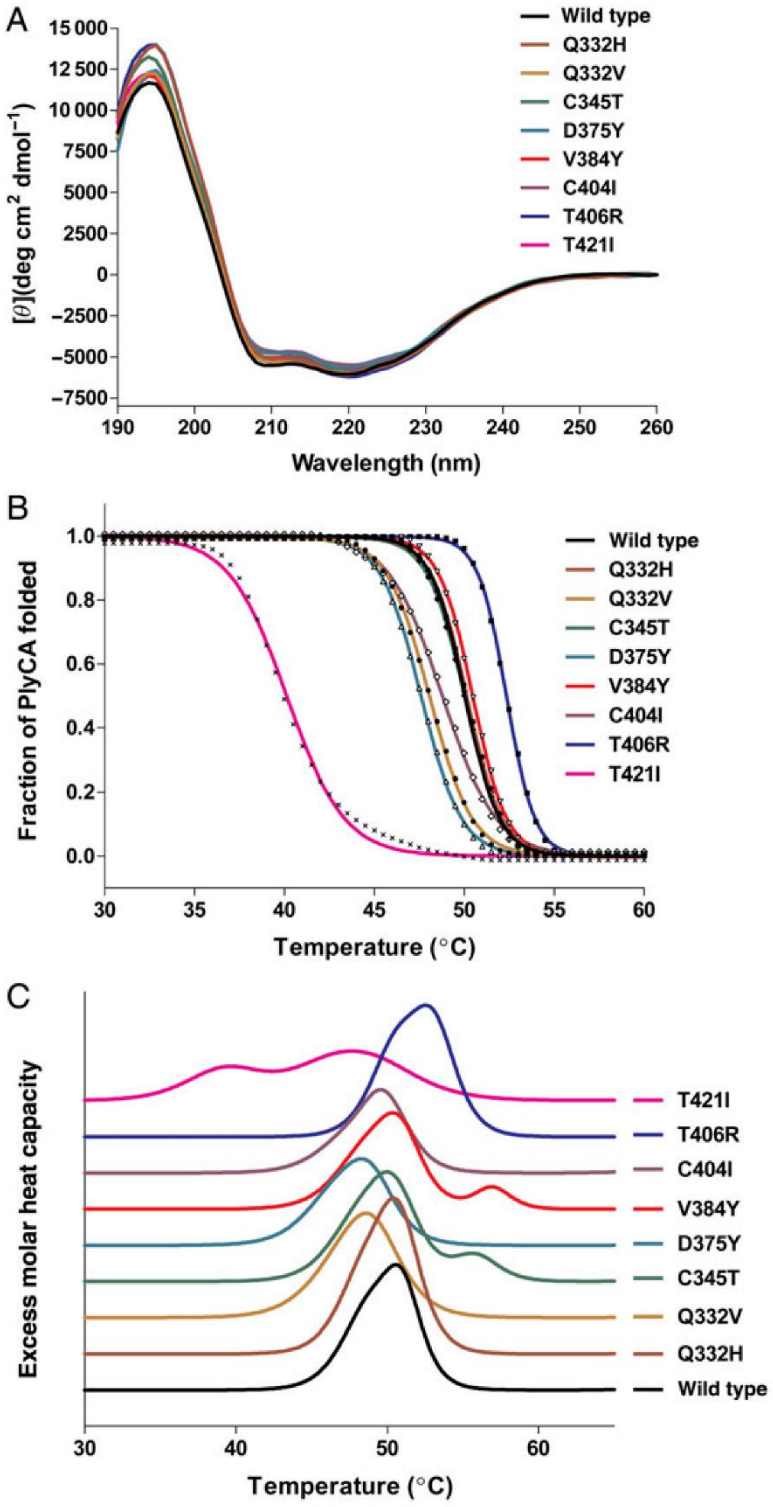


Figure 4-3. SDS-PAGE analysis of the various PlyC constructs experimentally characterized. The solubility and purity of each enzyme was analyzed on a 4-15% gradient SDS-PAGE gel. The various lanes correlate to: (M) Molecular weight standard, (1) Wild-type, (2) PlyC (PlyCA) Q332H, (3) Q332V, (4) C345T, (5) D375Y, (6) V384Y, (7) C404I, (8) T406R and (9) T421I. Protein expression was not observed for PlyC (PlyCA) D330Y and T381Y and therefore both were excluded from the gel.

Figure 4-4. Secondary structure and thermal stability determination. (a) The secondary structure of each of the PlyC construct was analyzed by far-UV CD spectroscopy from 190-260 nm. The mean residue ellipticity [θ] (deg cm^2 $dmol^{-1}$) was plotted against wavelength (nm) for each mutant, with all of the resulting spectra being overlaid for comparative purposes. The thermal stability of each mutant was then analyzed by means of (b) CD thermal denaturation and (c) DSC experiments in 20 mM phosphate buffer, pH 7.0, at a protein concentration of 1 mg/ml using a heating rate of 1 ℃/min. Note, only data for PlyCA is depicted for CD and DSC analysis as PlyCB denatures >70 ℃.



(Figure 4-4, See above for caption)

4.4.2 Kinetic Analysis of Bacteriolytic Activity against S. pyogenes

To assess the bacteriolytic activity of each PlyC mutant, the purified enzymes were incubated with *S. pyogenes* D471 at different molar concentrations and the resulting activity was elucidated by turbidity reduction assays. There were no activity defects observed with the PlyC (PlyCA) C345T, D375Y and V384Y mutants, displaying 1.13, 1.03 and 1.23 fold increases in activity when compared to wild-type, respectively (Table 4-2). PlyC (PlyCA) mutants C404I and T406R exhibited a moderate loss in activity, exhibiting a respective 2.2 and 2.1 fold decrease in activity. More significant activity deficiencies were observed with the PlyC (PlyCA) Q332H, Q332V and T421I mutants, of 3.6, 9.1 and 16.7 fold reduction in activity, respectively.

Table 4-2. Bacteriolytic activity quantitation by means of *S. pyogenes* turbidity reduction assay.

PlyC Construct	Lytic Activity (U/ml)	Relative Lytic Activity
Wild-type	45350	1.00
Q332H	12730	0.28
Q332V	4950	0.11
C345T	51180	1.13
D375Y	46550	1.03
V384Y	55560	1.23
C404I	16000	0.45
T406R	21550	0.48
T421I	2880	0.06

4.4.3 Circular Dichroism Thermal Stability Analysis

Equal molar concentrations of the PlyC mutant enzymes were subjected to CD thermal denaturation experiments to determine the T_G values of the mutagenized PlyCA subunits in the context of the holoenzyme structure (Figure 4-4b). Over the temperature range tested, distinct thermally-induced structural transitions were observed for both the PlyCA and PlyCB subunits for each mutant analyzed. When monitoring the loss of α-helical secondary structure, PlyCA qualitatively exhibits a single, cooperative structural transition that is not reversed on cooling. PlyC (PlyCA) mutations Q332V, C345T, D375Y, C404I and T421I were destabilizing, decreasing the T_G of the catalytic subunit by 1.98 ℃, 0.07 ℃, 2.45 ℃, 0.99 ℃ and 10.41 ℃, respectively, when compared to wild-type (T_G = 50.09 ℃) (Table 4-3). The PlyC (PlyCA) mutants Q332H and V384Y slightly stabilized PlyCA by 0.08 ℃ and 0.39 ℃, respectively. The T406R point mutation to the catalytic subunit of PlyC was the most structurally stabilizing, augmenting the T_G of PlyCA by 2.27 ℃.

Table 4-3. Biophysical thermal analysis of wild-type PlyC and the computationally predicted stabilizing point mutants. Results from CD thermal denaturation (columns 2 and 3) and DSC (columns 4-9) experiments are depicted for the PlyCA subunit only.

PlyC Construct	Circular Dichroism		Differential Scanning Calorimetry					
	T_G (°C)	ΔT_G (°C)	T_{G1} (°C)	T_{G2} (°C)	T_{G3} (°C)	ΔH_{VH1} (kcal/mol)	ΔH_{VH2} (kcal/mol)	ΔH_{VH3} (kcal/mol)
Wild-type	50.09	-	48.27	50.67		171.23	205.09	
Q332H	50.17	+0.08	48.47	50.74		177.07	211.87	
Q332V	48.11	-1.98	46.94	49.28		150.50	192.22	
C345T	50.02	-0.07	48.00	50.45	55.56	158.39	205.09	193.50
D375Y	47.64	-2.45	46.26	48.78		156.01	181.05	
V384Y	50.48	+0.39	48.50	50.79	57.05	173.96	212.65	252.90
C404I	49.10	-0.99	47.72	49.99		153.76	203.14	
T406R	52.36	+2.27	50.48	53.05		192.68	211.38	
T421I	39.68	-10.41	39.24	45.75	49.64	121.37	113.83	118.01

4.4.4 Differential Scanning Calorimetry

To validate the CD structural stability analysis, the thermal denaturation of each PlyC mutant was investigated by DSC at equal molar concentrations (Figure 4-4c). Thermal transitions corresponding to the unfolding of both the PlyCA and PlyCB components of the holoenzyme were observed for each mutant inspected. DSC analysis of the FoldX mutants depicts PlyCA unfolding to fulfill a three-state, and in some cases, four-state, thermal transition model. Heating the protein samples from 15 ℃ to 105 ℃ followed by immediate cooling from 105 ℃ to 15 ℃ did not result in the refolding of PlyCA or PlyCB for each PlyC construct investigated (data not shown).

Consistent with CD results, PlyC (PlyCA) mutants Q332V, C345T, D375Y, C404I and T421I were less thermostable than wild-type (T_G = 48.27 ℃, van't Hoff enthalpy of unfolding (ΔH_{VH}) = 171.23 kcal/mol) when analyzed by DSC, encompassing a 1.33 ℃, 0.27 ℃, 2.01 ℃, 0.52 ℃ and 9.03 ℃ decrease in PlyCA T_G and a 20.73 kcal/mol, 12.84 kcal/mol, 15.22 kcal/mol, 17.47 kcal/mol and 49.86 kcal/mol reduction in ΔH_{VH}, respectively (Table 4-3). PlyC (PlyCA) mutants Q332H and V384Y displayed marginable increases in stability, with an increase in PlyCA T_G of 0.20 ℃ and 0.23 ℃, and a 5.84 kcal/mol and 2.73 kcal/mol gain in ΔH_{VH} , respectively. The T406R mutation produces a more notable improvement in the thermal fitness of PlyCA, improving the T_G and ΔH_{VH} by 2.21 ℃ and a 21.45 kcal/mol, respectively.

4.4.5 45 ℃ Kinetic Inactivation Analysis

The rate of thermally-induced kinetic inactivation was monitored for wild-type PlyC and the lead FoldX mutant candidate, PlyC (PlyCA) T406R, at 45 ℃ for a total of 3 hours. For this particular assay, the unfolding of PlyCA is directly correlated with the loss of bacteriolytic activity as a function of temperature and time. The loss in activity is not associated with the unfolding of the PlyCB binding domain of PlyC due to the inherent thermal stability of the octameric CBD complex of the CBD. The heat-labile nature of wild-type PlyC promoted rapid PlyCA unfolding at 45 ℃, resulting in a half-life ($t_{1/2}$) of 17.84 min (Figure 4-5, squares). Conversely, PlyC (PlyCA) T406R mutant improved the kinetic fitness of the enzyme 16 fold at 45 ℃ when compared to

wild-type, displaying an extrapolated $t_{1/2}$ increase to 286.09 minutes (Figure 4-5, inverted triangles).

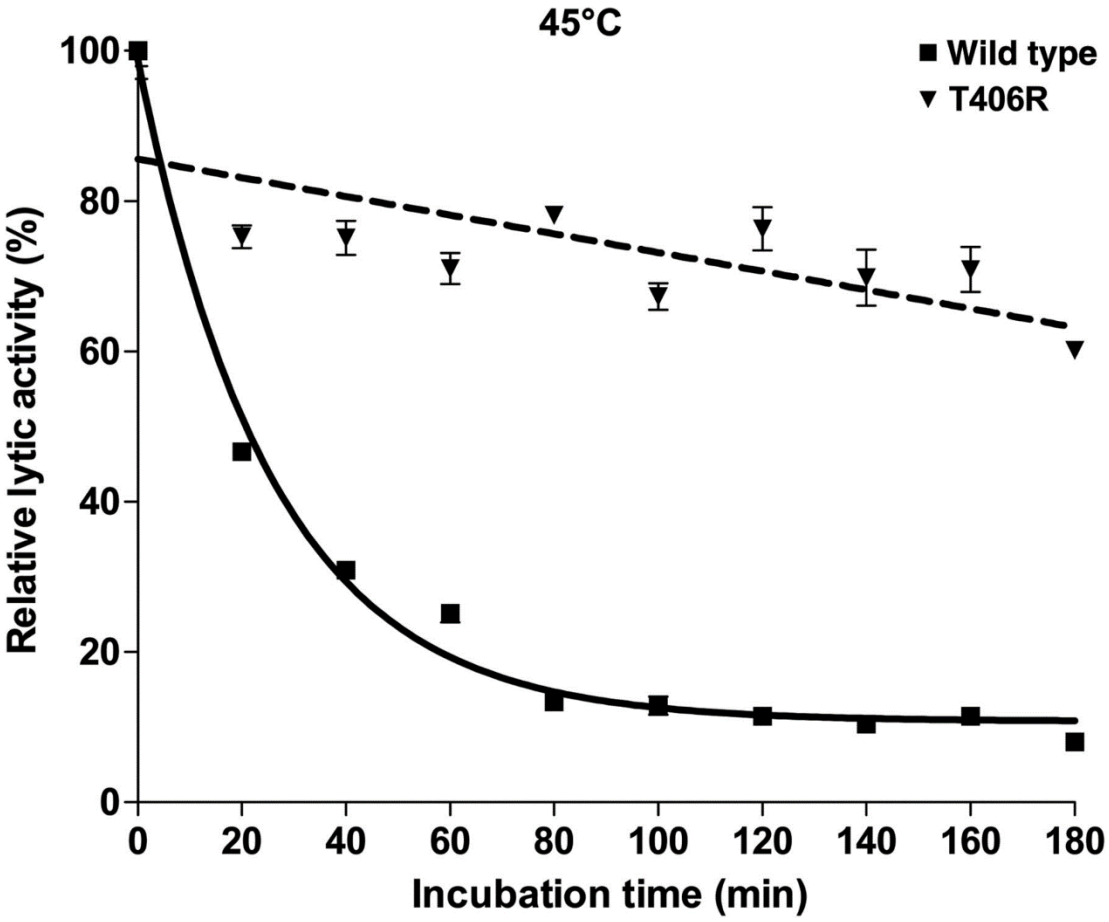


Figure 4-5. Kinetic stability of wild-type PlyC and PlyC (PlyCA) T406R at 45 ℃. Equal molar concentrations of wild-type PlyC and PlyC (PlyCA) T406R were incubated at 45 ℃ for a total of 3 hours. At 20 minute increments, the residual lytic activity of each enzyme was monitored by means of turbidity reduction assay. The activity of each was normalized to the lytic activity displayed by the unheated sample. All data were expressed as the mean ± SEM of triplicate experiments.

4.5 Discussion

Nineteen of the final 31 FoldX mutants were validated by Rosetta to be stabilizing, with the other 12 mutations calculated to be energetically unfavorable. Similar to FoldX, Rosetta appeared to favor mutations at the polar regions of the CHAP domain, with 12 of the predicted 19 advantageous point mutations being located at the surface. However, the percentage of the predicted stabilizing mutations located at the surface versus the hydrophobic core seems to be more evenly distributed for Rosetta (63% versus 37%) than FoldX (74% versus 26%). Additionally, the number of favorable mutations predicted to be near the active site was decidedly higher using FoldX (42%) than Rosetta (26%).

A final top ten candidate list, which was assembled based on the most favorable $\Delta\Delta G$ values independently validated by the FoldX and Rosetta algorithms, was experimentally analyzed. Due to an absence of protein expression, two of the ten candidates were immediately eliminated from further characterization. Of the eight experimentally characterized PlyC point mutants, only three of the mutations were found to be thermostabilizing (Figure 4-4b and c, Table 4-3). Both of the mutations to the hydrophobic core region of the CHAP domain were unfavorable, while three of the six surface mutations increased the stability of PlyCA. Of the three stabilizing mutations, the two mutations distant from the active site of CHAP gave the largest improvements in thermal stability. There was no correlation between the magnitude of the $\Delta\Delta G$ values estimated by the FoldX and Rosetta algorithms and the calorimetrically measured stability of the eight PlyC point mutants. For example, the

mutant that exhibited the greatest gain in stability, PlyC (PlyCA) T406R, was predicted to be the second least favorable mutation of the eight experimentally tested by both FoldX and Rosetta. Conversely, despite possessing the second most stabilizing $\Delta\Delta G$ value calculated by both algorithms, the PlyC (PlyCA) T421I mutant displayed the highest degree of thermolability of the eight mutants thermally characterized.

With respect to the effectiveness of the FoldX and Rosetta algorithms for estimating CHAP domain mutant $\Delta\Delta G$ values, it should be born in mind that these calculations are based on a relatively low-resolution 3.3-Å X-ray crystal structure that is highly complex and dynamic. Computationally modeling mutations into an intricate nine-subunit holoenzyme structure with incomplete atomic coordinates and a flexible catalytic subunit could contribute to the inconsistency between the computational and experimental data. In addition, there are many approximations in the computational methods, and neither adequately treats contributions from altered dynamics resulting from the mutations. Thus the low correlation between predicted and observed effects on thermostability is not surprising. Nevertheless, the $\Delta\Delta G$ estimates derived from FoldX ultimately did yield one very useful and non-obvious candidate that increased the stability of PlyCA, at the expense of some extra experimental work on non-useful ones. There may, of course, be other potentially useful mutations that the procedure used here overlooked.

Although the structural integrity of all eight of the remaining PlyC mutants remained intact (Figure 4-3 and Figure 4-4a), endolysin turbidity reduction activity titers showed that while the C404I and T406R mutations respectively caused considerable 2.2 and 2.1 fold decreases in activity, the Q332H, Q332V and T421I mutations to the CHAP domain generated significant activity defects that correlated to 3.6-16.7 fold losses in activity (Table 4-2). Considering the PlyCA CHAP domain has active site residues at C333 and H420 (Nelson et al., 2006), it was not surprising to observe major perturbations to the catalytic efficiency of the enzyme when introducing amino acid mutations adjacent to either of the two active site residues. Of the five point mutations that conferred a loss in bacteriolytic activity, four of these mutations were located in the hydrophobic core and/or near the active site of the CHAP domain. There is no correlation between the activity displayed by a particular PlyC construct and the extent of its predicted $\Delta\Delta G$ values by either FoldX or Rosetta.

After being subjected to a biophysical thermal analysis, the lead mutant candidate was PlyC (PlyCA) T406R, which displayed a ~2.2 ℃ increase in PlyCA thermal denaturation temperature with a 21.45 kcal/mol gain in ΔH_{VH} (Figure 4-4b and 4-4c, Table 4-3). The T406R mutation is located on the CHAP domain surface and is hypothesized to promote an inter-domain interaction between the N- and C-terminal domains of the PlyCA subunit. Modeling the T406R mutation into the CHAP domain shows how the elongated arginine side-chain allows the formation of a stabilizing hydrogen bond with the polar Q106 side-chain located on the surface of the N-terminal GyH domain of PlyCA (Figure 4-6). The suspected PlyCA stabilizing inter-

domain interaction engineered by the T406R mutation also resulted in an increase in kinetic fitness, with the point mutant promoting an extrapolated 16 fold augmentation in kinetic stability at 45 ℃ (Figure 4-5). Although PlyC (PlyCA) T406R had an overall reduction in bacteriolytic activity when compared to the endogenous activity of wild-type PlyC, a common observation when thermostabilizing biomolecules (Arnold, Wintrode, Miyazaki, & Gershenson, 2001; Beadle & Shoichet, 2002; Giver, Gershenson, Freskgard, & Arnold, 1998; Meiering, Serrano, & Fersht, 1992; Mukaiyama et al., 2006; Shoichet, Baase, Kuroki, & Matthews, 1995; Yutani, Ogasahara, Tsujita, & Sugino, 1987), the residual activity displayed by the mutant nonetheless remains more potent than that of any other characterized endolysin.

It is important to keep in mind that, although the ~2.2 ℃ increase in thermal denaturation temperature displayed by the PlyC (PlyCA) T406R mutant was modest, this mutation provoked a pronounced 16 fold improvement in kinetic stability. Moreover, engineering significantly enhanced thermal stability to proteins is generally achieved through combining multiple thermostabilizing amino acid mutations that individually have a small effect on stability (Akasako, Haruki, Oobatake, & Kanaya, 1995; Ohage & Steipe, 1999; Pantoliano et al., 1989; Serrano, Day, & Fersht, 1993; Shih & Kirsch, 1995; von der Osten et al., 1993). If the mechanism of stabilization employed by each individual mutation is unique, then combining these advantageous mutations can additively stabilize the protein. To this end, the thermally advantageous mutations identified in this computational study

(Q332H, V384Y, T406R) could be combined to possibly further progress the thermal properties of PlyC.

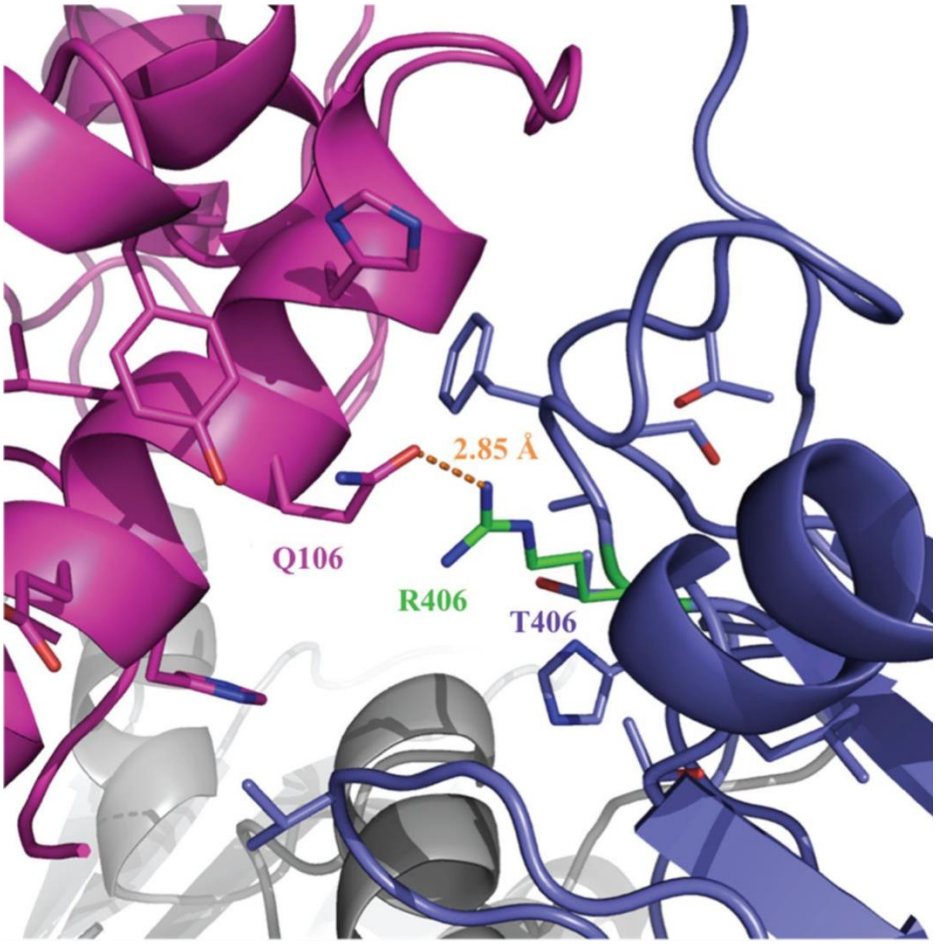


Figure 4-6. Local structure around wild-type PlyCA T406 with the proposed conformation of the mutant T406R superimposed. The crystal structure of the wild-type PlyCA T406 residue (blue sticks) and the model of the PlyCA mutant residue T406R (green sticks) are shown together with the surrounding residues. The predicted additional hydrogen bond between Q106 of the N-terminal GyH domain and R406 of the C-terminal CHAP domain is shown as orange dots. Parts of the polypeptide backbone of the PlyCA N-terminal GyH domain (magenta) and the C-terminal CHAP domain (blue) are also shown.

4.6 Acknowledgements

We would like to respectively thank Amanda Altieri and Robert G. Brinson for their technical assistance when using the CD spectrometer and DSC. We are also thankful for the insight and advice provided by Philip N. Bryan regarding the thermodynamic and kinetic analysis experiments. Finally, we are grateful to the National Institute of Standards and Technology for access to DSC instrumentation.

Author contributions: R.D.H, J.M. and D.C.N. conceived the experiments; Y.Y. and J.M. performed the computational experiments and analyzed the data; R.D.H. performed the biological, biochemical and biophysical experiments and analyzed the data; J.M and D.C.N. planned and supervised the project. R.D.H. wrote the article.

Funding: This work was supported by grants from the United States Department of Defense (DM102823) and the Maryland Agriculture Experiment Station to DCN.

Chapter 5: Conclusion and perspectives

In this dissertation, I used various computational methods to analyze and predict the molecular, phenotypic, and pathogenic effects of missense mutations. In the last chapter, I briefly summarize the conclusions of each project and look to the future in these fields.

5.1 Brief summary

In the first project, I developed a new ensemble approach to address a largely unsolved mutation interpretation problem – predicting continuous phenotype values, in one case for the enzyme activity of a set of rare human mutations in a monogenic disease gene and in the other for a yeast complementation growth assay for mutations of human SUMO-ligase. The ensemble approach was relatively effective for this task, as well as for regular binary pathogenicity assignments. In addition, I investigated the potential of the ensemble method in estimating the reliability of pathogenicity assignments for better clinical applicability. Next, I characterized and compared the mutations in monogenic disease and in cancer, looking for the unique features of cancer driver mutations and directions to improve current mutation interpretation methods on cancer data. The results pinpointed the issue of passenger mutations in cancer driver genes and confirmed the applicability of general interpretation methods and properties of mutations for three structure related protein features. Finally, I conducted a protein thermostability engineering study by computationally interpreting

mutation effects on protein stability. Experiments by our collaborators showed encouraging success, although significant improvements can be made in several aspects of that task, as discussed below.

5.2 The demand for the right dataset

Methods to interpret mutation effects in monogenic diseases and complex trait diseases often rely on training datasets that comprises a case set of disease-related mutations collected from literature or curated databases such as HGMD (Stenson et al., 2014), UniProt (UniProt consortium 2015), OMIM (http://omim.org/), and ClinVar (Landrum et al., 2016), and a control set of neutral mutations obtained from observed variants across species or polymorphisms in populations. The computational methods developed based on these datasets are mostly designed to make a binary assignment of pathogenic or nonpathogenic, and have been benchmarked in many studies (Dong et al., 2015; Gnad et al., 2013; Thusberg, Olatubosun, & Vihinen, 2011). It was less clear how these methods perform on more realistic datasets such as in the CAGI NAGLU challenge (Hoskins et al., 2017), where mutation effects are distributed more evenly across the entire range of functional activity. Mutations with a mid-range molecular or functional effects still pose a challenge for all contemporary approaches. To fully investigate this issue, there is a major requirement for more realistic training and testing datasets like NAGLU. However, one impediment is to find an appropriate way to experimentally assay the target proteins and to scale up the study.

The advent of high-throughput techniques such as deep mutational scanning in the last several years provides a potential solution to this. Like in the CAGI SUMO-ligase challenge, these techniques are able to measure the functional consequences of thousands of mutants, together with massive sequencing to identify corresponding mutations (Fowler & Fields, 2014; Wrenbeck, Faber, & Whitehead, 2017). On the other hand, there are inherent limitations to these techniques. I investigated the performance of an ensemble method on a set of deep mutational scanning datasets and found poorer results than on more traditional single measurement datasets. Potential issues include data quality and stochastic error (Fowler & Fields, 2014). This issue needs to be addressed more thoroughly and rigorously, for example with the statistical framework proposed recently (Rubin et al., 2017). Moreover, deep mutational scans are usually implemented in the growth-based assay, phage display or cell flow sorting, which are only capable of measuring a limited set of phenotypes. In the SUMO-ligase challenge, another possible complication is that the interfaces between human SUMO-ligase and its binding partners in human and in yeast may have different properties. Nevertheless, the high-throughput techniques will help provide more valuable data for mutation effect analysis in the future.

In both the NAGLU and SUMO-ligase challenges, a potential cause of apparent false positive predictions arises from the use of cDNA constructs in the experiments. As a result, those missense mutations that affect splicing or alter the ratio of alternatively spliced isoforms *in vivo* (D'Souza et al., 1999; Ward & Cooper, 2009) would not affect the experimental phenotypes. For the NAGLU challenge, I checked for the

overlap of missense mutations with bases critical to splicing and found none. However, effects on splice enhancers and silencers (D'Souza et al., 1999; Ward & Cooper, 2009) are less straightforward to detect, and these may contribute false positives.

In the analyses of both monogenic disease and cancer, the negative control dataset was compiled using interspecies variants, that is amino acid differences in other species compared to human, assuming these substitutions would be benign in the human protein. However, it has been estimated that, on average, around 10% of these interspecies variants could be pathogenic to humans, but benign to other species due to compensation by substitutions at other sites (Kondrashov, Sunyaev, & Kondrashov, 2002). This issue was reduced in our analysis by excluding known human pathogenic mutations from the observed interspecies variants. A more sophisticated approach would be to examine all coevolving sites in the sequence profile of the homolog proteins and their potential interacting partners.

In the past decade, cancer research has greatly benefited from the explosion of cancer whole genome and exome sequencing data. The accuracy of many driver gene predictions depends on these large-scale data. On the other hand, interpreting and identifying the few cancer driver mutations in individuals creates a demand for a large gold standard dataset of the experimentally verified driver and passenger mutations. Notable community efforts have been made to address this issue, such as

the ICGC's Pan-Cancer Analysis of Whole Genomes (PCAWG) (https://dcc.icgc.org/pcawg).

For both complex trait disease and cancer, there is also a significant contribution from non-coding mutations. Generating large datasets in this regard is a new hotspot of research, and has been undertaken with great energy (GTEx Consortium et al., 2017; X. Li et al., 2017). New methods utilizing these datasets are now under development.

5.3 Improving mutation interpreting methods

A lesson learned from the CAGI NAGLU and SUMO-ligase challenges is that no contemporary mutation interpretation method provides revolutionary accuracy in predicting the experimental activity values. This suggests inherent deficiencies in the current prediction models that primarily relate sequence conservation patterns to pathogenicity. There are a number of ways in which more realistic evolutionary information could be utilized to provide potential improvements. These include the better utilization of phylogenetic information and incorporation of the relative likelihood of particular mutation types such single versus double base changes, transition/transversion relative frequencies and CpG island hotspots.

As noted earlier, most current methods produce binary predictions of pathogenic or non-pathogenic, whereas prediction of phenotypic variables on a continuous scale is

desirable. My ensemble approach, combining confidence scores for binary methods, provides one solution to this problem. It would also be interesting to explore methods that directly predict a continuous value. One approach would be to use appropriate machine learning methods, such as support vector regression.

A few contemporary methods also partially or fully rely on protein structure information. These methods can be improved by including state of the art structure modeling methods that take into account backbone flexibility and extensive rotamer optimization, such as ROSETTA (B Kuhlman & Baker, 2000; Rohl et al., 2004). Current models mainly focus on the destabilizing effect of mutations, which plays an important role in human diseases (Casadio, Vassura, Tiwari, Fariselli, & Luigi Martelli, 2011; Redler et al., 2016; Shi & Moult, 2011; Yue et al., 2005). However, the effects of some mutations, especially in cancer oncogenes where more driver mutations are located on the protein surface, manifest under various mechanisms such as catalytic activity, specificity, binding affinity and protein flexibility. While modeling on some of these are still challenging, molecular dynamics techniques provide a potential solution to tackle the modeling problems such as in intrinsically disordered regions and at the protein-macromolecular interfaces (Agarwal, Annamalai, Maiti, & Arsad, 2016; Doss, Chakraborty, Chen, & Zhu, 2014; George Priya Doss et al., 2014).

Additional improvement of methods may also come from 1) differentially treating mutations subsets, such as surface and core mutations, 2) retraining models on the mutations where methods do not correlate well, 3) developing gene-specific models, and 4) better frameworks that integrate sequence conservation-based methods with structure-based methods.

5.4 Bridging the gap between mutation research and clinical application

In the first project of this dissertation, I investigated the potential of ensemble methods to estimate the reliability for pathogenicity assignments, which is important in clinical applications because an accurate method may still fail on a certain subset of mutations. So far, this issue has not been well investigated. As a consequent, present clinical guidelines treat computational interpretation of potential disease mutations as only secondary evidence of a genetic cause (Richards et al., 2015). The ensemble methods achieved a substantial fraction (up to 40%) of pathogenicity assignments with clinically meaningful confidence (>90%). Future work and more testing on more data in various diseases will help increase the clinical applicability.

A large set of cancer driver genes have been prioritized from examination of large-scale cancer whole genome and exome data. A clinically more relevant task is to identify all driver mutations given a single cancer genome or exome. This creates a demand for methods to identify a long tail of potential driver genes with rare driver mutations. Moreover, it would be beneficial to establish models accounting for a

continuous cancer driver mutation penetrance instead of the current binary ‘driver vs. passenger’ model.

5.5 Beyond simple approaches for protein thermostability engineering

The T406R mutation we identified in the PlyCA CHAP domain (Heselpoth, Yin, Moult, & Nelson, 2015) displayed limited improvement in thermal stability, typical of that achievable with single mutations. Achieving a large improvement in stability usually requires a combination of multiple stabilizing mutations that operate in an additive or non-addition manner. On the other hand, the low success rate (3 out of 10) of our method reflects the serious false positive problem. Potential causes include but are not limited to the low-quality modeling from template structures, use of a rigid backbone protocol in structure modeling, and not including more structure modeling methods that can provide a consensus set of candidates. A recent work (Goldenzweig et al., 2016) reported a significant success in engineering thermally stable mutants (a 20°C increase in denaturation temperature), based on a strategy that combined searching for multiple stabilizing mutations with a method to remove false positives using phylogeny information. In future, one would expect this type of strategy to be more common. Indeed, our collaborator found a higher thermostability when combined T406R with another stabilizing mutation previously identified through a directed evolution approach. Computational methods also have the potential to engineer for thermostability in a broader range of conditions, such as altered pH (Strauch et al., 2014) and protein concentration (Goldenzweig et al., 2016).

Another issue with the successful T406R mutation was that it has a cost of more than half catalytic activity, as observed for many stabilizing mutations (Arnold et al., 2001; Beadle & Shoichet, 2002; Mukaiyama et al., 2006). It is still challenging to rationally model mutation effects in the context of enzyme activity that involves complicated physical-chemical calculation of the transition state complex. Enzyme activity may also be affected by changes in protein flexibility and dynamics. Such issues can be addressed by better integrating results of molecular dynamics methods and normal mode analysis.

Appendix A

Full list of my publications

A new method for disease missense mutation analysis:

Yin Y, Kundu K, Pal LR, Moult J. 2017. Ensemble variant interpretation methods to predict enzyme activity and assign pathogenicity in the CAGI4 NAGLU (Human N-acetyl-glucosaminidase) and UBE2I (Human SUMO-ligase) challenges. *Human Mutation.* 38(9):1109-1122.

Disease missense mutation analysis as a part of teamwork related to CAGI:

Kundu K, Pal LR, **Yin Y**, Moult J. 2017. Determination of disease phenotypes and pathogenic variants from exome sequence data in the CAGI 4 gene panel challenge. *Human Mutation.* 38(9):1201-1216.

Pal LR, Kundu K, **Yin Y**, Moult J. 2017. CAGI4 SickKids clinical genomes challenge: A pipeline for identifying pathogenic variants. *Human Mutation.* 38(9):1169-1181.

Pal LR, Kundu K, **Yin Y**, Moult J. 2017. CAGI4 Crohn's exome challenge: Marker SNP versus exome variant models for assigning risk of Crohn disease. *Human Mutation.* 38(9):1225-1234.

CAGI challenge assessor papers including my work:

Carraro M, Minervini G, Giollo M, Bromberg Y, Capriotti E, Casadio R, Dunbrack R, Elefanti L, Fariselli P, Ferrari C, Gough J, Katsonis P, Leonardi E, Lichtarge O, Menin C, Martelli PL, Niroula A, Pal LR, Repo S, Scaini MC, Vihinen M, Wei Q, Xu Q, Yang Y, **Yin Y**, Zaucha J, Zhao H, Zhou Y, Brenner SE, Moult J, Tosatto SCE. 2017. Performance of in silico tools for the evaluation of p16INK4a (CDKN2A) variants in CAGI. *Human Mutation*. 38(9):1042-1050.

Chandonia JM, Adhikari A, Carraro M, Chhibber A, Cutting GR, Fu Y, Gasparini A, Jones DT, Kramer A, Kundu K, Lam HYK, Leonardi E, Moult J, Pal LR, Searls DB, Shah S, Sunyaev S, Tosatto SCE, **Yin Y**, Buckley BA. 2017. Lessons from the CAGI-4 Hopkins clinical panel challenge. *Human Mutation*. 38(9):1155-1168.

Cai B, Li B, Kiga N, Thusberg J, Bergquist T, Chen YC, Niknafs N, Carter H, Tokheim C, Beleva-Guthrie V, Douville C, Bhattacharya R, Yeo HTG, Fan J, Sengupta S, Kim D, Cline M, Turner T, Diekhans M, Zaucha J, Pal LR, Cao C, Yu CH, **Yin Y**, Carraro M, Giollo M, Ferrari C, Leonardi E, Tosatto SCE, Bobe J, Ball M, Hoskins RA, Repo S, Church G, Brenner SE, Moult J, Gough J, Stanke M, Karchin R, Mooney SD. 2017. Matching phenotypes to whole genomes: Lessons learned from four iterations of the personal genome project community challenges. *Human Mutation*. 38(9):1266-1276.

Daneshjou R, Wang Y, Bromberg Y, Bovo S, Martelli PL, Babbi G, Lena PD, Casadio R, Edwards M, Gifford D, Jones DT, Sundaram L, Bhat RR, Li X, Pal LR, Kundu K, **Yin Y**, Moult J, Jiang Y, Pejaver V, Pagel KA, Li B, Mooney SD, Radivojac P, Shah S, Carraro M, Gasparini A, Leonardi E, Giollo M, Ferrari C, Tosatto SCE, Bachar E, Azaria JR, Ofran Y, Unger R, Niroula A, Vihinen M, Chang B, Wang MH, Franke A, Petersen BS, Pirooznia M, Zandi P, McCombie R, Potash JB, Altman RB, Klein TE, Hoskins RA, Repo S, Brenner SE, Morgan AA. 2017. Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. *Human Mutation*. 2017 Sep;38(9):1182-1192.

<u>Projects on protein design and modeling:</u>

Heselpoth RD, **Yin Y**, Moult J, Nelson DC. 2015. Increasing the stability of the bacteriophage endolysin PlyC using rationale-based FoldX computational modeling. *Protein Engineering Design and Selection.* 28(4):85-92.

Shen Y, Barros M, Vennemann T, Gallagher DT, **Yin Y**, Linden SB, Heselpoth RD, Spencer DJ, Donovan DM, Moult J, Fischetti VA, Heinrich F, Lösche M, Nelson DC. 2016. A bacteriophage endolysin that eliminates intracellular streptococci. *Elife*. e13152.

Appendix B

UNIVERSITY OF MARYLAND
BIOLOGICAL SCIENCES GRADUATE PROGRAM

2101 Bioscience Research Building
College Park, Maryland 20742-4415
301.405.6905/6991 TEL, 301.314.9921 FAX

The Graduate School
2123 Lee Building
University of Maryland
College Park, MD 20742

This letter is written to signify that the dissertation committee, committee chair, and the graduate director have all approved the use of previously published co-authored work in the final dissertation of Yizhou Yin, Biological Sciences Graduate Program, 109077491. In accordance with the Graduate School's policy the dissertation committee has determined that they made substantial contributions to the included work.

The citations for the published work are:

Yin Y, Kundu K, Pal LR, Moult J. 2017. Ensemble variant interpretation methods to predict enzyme activity and assign pathogenicity in the CAGI4 NAGLU (Human N-acetyl-glucosaminidase) and UBE2I (Human SUMO-ligase) challenges. *Human Mutation*. 38(9):1109-1122.

Heselpoth RD, Yin Y, Moult J, Nelson DC. 2015. Increasing the stability of the bacteriophage endolysin PlyC using rationale-based FoldX computational modeling. *Protein Engineering Design and Selection*. 28(4):85-92.

Per Graduate School policy the dissertation forward will identify the scope and nature of the student's contributions to the jointly authored work included in the dissertation and a copy of this letter will be submitted with the dissertation.

Sincerely,

11/6/17

John Moult, Dissertation Committee Chair,
Professor, Department of Cell Biology and Molecular Genetics

Dr. Michelle Brooks,
Associate Director, Biological Sciences Graduate Program

Yizhou Yin, 11/06/2017
Graduate Student, Biological Sciences

Bibliography

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Min Kang, H., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74.

Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., & Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics (Oxford, England)*, *26*(3), 392–8. https://doi.org/10.1093/bioinformatics/btp630

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249.

Agarwal, T., Annamalai, N., Maiti, T. K., & Arsad, H. (2016). Biophysical changes of ATP binding pocket may explain loss of kinase activity in mutant DAPK3 in cancer: A molecular dynamic simulation analysis. *Gene*, *580*(1), 17–25.

Akasako, A., Haruki, M., Oobatake, M., & Kanaya, S. (1995). High resistance of Escherichia coli ribonuclease HI variant with quintuple thermostabilizing mutations to thermal denaturation, acid denaturation, and proteolytic degradation. *Biochemistry*, *34*(25), 8115–8122.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–421. https://doi.org/10.1038/nature12477

Alzheimer's Association. (2015). 2015 Alzheimer's disease facts and figures. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *11*(3),

332–384.

Anderson, G., & Scott, M. (1991). Determination of product shelf life and activation energy for five drugs of abuse. *Clinical Chemistry*, *37*(3), 398–402.

Armitage, P., & Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer*, *8*(1), 1–12.

Arnold, F. H., Wintrode, P. L., Miyazaki, K., & Gershenson, A. (2001). How enzymes adapt: lessons from directed evolution. *Trends in Biochemical Sciences*, *26*(2), 100–106.

Ashworth, J., Havranek, J. J., Duarte, C. M., Sussman, D., Monnat Jr., R. J., Stoddard, B. L., & Baker, D. (2006). Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature*, *441*(7093), 656–659.

Baird, P. A., Anderson, T. W., Newcombe, H. B., & Lowry, R. B. (1988). Genetic disorders in children and young adults: a population study. *American Journal of Human Genetics*, *42*(5), 677–693.

Baugh, E. H., Simmons-Edler, R., Müller, C. L., Alford, R. F., Volfovsky, N., Lash, A. E., & Bonneau, R. (2016). Robust classification of protein variation using structural modelling and large-scale data integration. *Nucleic Acids Research*, *44*(6), 2501–13. https://doi.org/10.1093/nar/gkw120

Beadle, B. M., & Shoichet, B. K. (2002). Structural bases of stability-function tradeoffs in enzymes. *Journal of Molecular Biology*, *321*(2), 285–296.

Beesley, C., Moraitou, M., Winchester, B., Schulpis, K., Dimitriou, E., & Michelakakis, H. (2004). Sanfilippo B syndrome: molecular defects in Greek patients. *Clinical Genetics*, *65*(2), 143–149. https://doi.org/10.1111/j.0009-

9163.2004.00210.x

Bernier-Villamor, V., Sampson, D. A., Matunis, M. J., & Lima, C. D. (2002). Structural basis for E2-mediated SUMO conjugation revealed by a complex between ubiquitin-conjugating enzyme Ubc9 and RanGAP1. *Cell*, *108*(3), 345–56. https://doi.org/10.1016/S0092-8674(02)00630-X

Blume-Jensen, P., & Hunter, T. (2001). Oncogenic kinase signalling. *Nature*, *411*(6835), 355–365.

Boycott, K. M., Vanstone, M. R., Bulman, D. E., & MacKenzie, A. E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics*, *14*(10), 681–691.

Bustamante, N., Rico-Lastres, P., Garcia, E., Garcia, P., & Menendez, M. (2012). Thermal stability of Cpl-7 endolysin from the Streptococcus pneumoniae bacteriophage Cp-7; cell wall-targeting of its CW_7 motifs. *PLoS One*, *7*(10), e46654.

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L., & Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. *Human Mutation*, *30*(8), 1237–1244.

Cancer Genome Atlas Network, Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., … Thomson., E. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, *487*(7407), 330–337.

Cancer Genome Atlas Network, Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., … Palchik, J. D. (2012). Comprehensive

molecular portraits of human breast tumours. *Nature*, *490*(7418), 61–70.

Cancer Genome Atlas Research Network, Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D. W., … Thomson, E. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, *474*(7353), 609–615.

Cancer Genome Atlas Research Network, Getz, G., Gabriel, S. B., Cibulskis, K., Lander, E., Sivachenko, A., … Levine, D. A. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, *497*(7447), 67–73.

Cancer Genome Atlas Research Network, Hammerman, P. S., Lawrence, M. S., Voet, D., Jing, R., Cibulskis, K., … Meyerson, M. (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, *489*(7417), 519–525.

Cancer Genome Atlas Research Network, Ley, T. J., Miller, C., Ding, L., Raphael, B. J., Mungall, A. J., … Eley, G. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *The New England Journal of Medicine*, *368*(22), 2059–2074.

Cancer Genome Atlas Research Network, McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., … Thomson, E. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, *455*(7216), 1061–1068.

Canutescu, A. A., Shelenkov, A. A., & Dunbrack, R. L. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Science*, *12*(9), 2001–2014.

Capriotti, E., Altman, R. B., & Bromberg, Y. (2013). Collective judgment predicts disease-associated single nucleotide variants. *BMC Genomics*, *14*(Suppl 3), S2.

https://doi.org/10.1186/1471-2164-14-S3-S2

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V. E., Kinzler, K. W., … Karchin, R. (2009). Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Research*, *69*(16), 6660–6667.

Carter, H., Douville, C., Stenson, P. D., Cooper, D. N., & Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics*, *14 Suppl 3*, S3.

Casadio, R., Vassura, M., Tiwari, S., Fariselli, P., & Luigi Martelli, P. (2011). Correlating disease-related mutations to their effect on protein stability: a large-scale analysis of the human proteome. *Human Mutation*, *32*(10), 1161–1170.

Cheng, J., Randall, A., & Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*, *62*(4), 1125–1132.

Chimpanzee Sequencing and Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, *437*(7055), 69–87.

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS One*, *7*(10), e46688.

Chun, S., & Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. *Genome Research*, *19*(9), 1553–1561.

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., & Sander, C.

(2013). Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, *45*(10), 1127–1133.

Cooper, G. M., & Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews. Genetics*, *12*(9), 628–640.

Cooper, G. M., Stone, E. A., Asimenos, G., NISC Comparative Sequencing Program, Green, E. D., Batzoglou, S., & Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, *15*(7), 901–913.

Correia, B. E., Bates, J. T., Loomis, R. J., Baneyx, G., Carrico, C., Jardine, J. G., … Schief, W. R. (2014). Proof of principle for epitope-focused vaccine design. *Nature*, *507*(7491), 201–216.

D'Souza, I., Poorkaj, P., Hong, M., Nochlin, D., Lee, V. M., Bird, T. D., & Schellenberg, G. D. (1999). Missense and silent tau gene mutations cause frontotemporal dementia with parkinsonism-chromosome 17 type, by affecting multiple alternative RNA splicing regulatory elements. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(10), 5598–5603.

Das, R., & Baker, D. (2008). Macromolecular modeling with rosetta. *Annual Review of Biochemistry*, *77*(1), 363–382.

Daude, D., Topham, C. M., Remaud-Simeon, M., & Andre, I. (2013). Probing impact of active site residue mutations on stability and activity of Neisseria polysaccharea amylosucrase. *Protein Science*, *22*(12), 1754–1765.

Davoli, T., Xu, A. W., Mengwasser, K. E., Sack, L. M., Yoon, J. C., Park, P. J., &

Elledge, S. J. (2013). Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, *155*(4), 948–962.

de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., … Barrett, J. C. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics*, *49*(2), 256–261.

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., … Ding, L. (2012). MuSiC: identifying mutational significance in cancer genomes. *Genome Research*, *22*(8), 1589–1598.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, *44*(3), 837. https://doi.org/10.2307/2531595

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning (pp. 1–15). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45014-9_1

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., & Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics*, *24*(8), 2125–2137.

Doss, C. G. P., Chakraborty, C., Chen, L., & Zhu, H. (2014). Integrating in silico prediction methods, molecular docking, and molecular dynamics simulation to predict the impact of ALK missense mutations in structural perspective. *BioMed Research International*, 895831.

Douville, C., Masica, D. L., Stenson, P. D., Cooper, D. N., Gygax, D. M., Kim, R., … Karchin, R. (2016). Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). *Human Mutation*, *37*(1), 28–35.

Drummond, D. A., Raval, A., & Wilke, C. O. (2006). A single determinant dominates the rate of yeast protein evolution. *Molecular Biology and Evolution*, *23*(2), 327–337.

Drummond, D. A., & Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, *134*(2), 341–352.

Dunbrack, R. L. J. (2002). Rotamer libraries in the 21st century. *Current Opinion in Structural Biology*, *12*(4), 431–440.

Eisenhaber, F., & Argos, P. (1993). Improved strategy in analytic surface calculation for molecular systems: Handling of singularities and computational efficiency. *Journal of Computational Chemistry*, *14*(11), 1272–1280.

Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C., & Scharf, M. (1995). The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *Journal of Computational Chemistry*, *16*(3), 273–284.

Fallas, J. A., & Hartgerink, J. D. (2012). Computational design of self-assembling register-specific collagen heterotrimers. *Nature Communications*, *3*, 1087.

Filatova, L. Y., Becker, S. C., Donovan, D. M., Gladilin, A. K., & Klyachko, N. L. (2010). LysK, the enzyme lysing Staphylococcus aureus cells: specific kinetic features and approaches towards stabilization. *Biochimie*, *92*(5), 507–513.

Fischetti, V. A., Nelson, D., & Schuch, R. (2006). Reinventing phage therapy: are the parts greater than the sum? *Nature Biotechnology*, *24*(12), 1508–1511.

Folkman, L., Stantic, B., Sattar, A., & Zhou, Y. (2016). EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models. *Journal of Molecular Biology*, *428*(6), 1394–1405. https://doi.org/10.1016/j.jmb.2016.01.012

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., … Campbell, P. J. (2017). COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research*, *45*(D1), D777–D783.

Fornili, A., Pandini, A., Lu, H.-C., & Fraternali, F. (2013). Specialized Dynamical Properties of Promiscuous Residues Revealed by Simulated Conformational Ensembles. *Journal of Chemical Theory and Computation*, *9*(11), 5127–5147.

Fowler, D. M., & Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nature Methods*, *11*(8), 801–807.

Francioli, L. C., Polak, P. P., Koren, A., Menelaou, A., Chun, S., Renkens, I., … Sunyaev, S. R. (2015). Genome-wide patterns and properties of de novo mutations in humans. *Nature Genetics*, *47*(7), 822–826.

Frank, E., Hall, M. A., & Witten, I. H. (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"* (4th ed.). Morgan Kaufmann.

Frishman, D., & Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Genetics*, *23*(4), 566–579.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., …

Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews. Cancer*, *4*(3), 177–183.

Gareau, J. R., Reverter, D., & Lima, C. D. (2012). Determinants of small ubiquitin-like modifier 1 (SUMO1) protein specificity, E3 ligase, and SUMO-RanGAP1 binding activities of nucleoporin RanBP2. *The Journal of Biological Chemistry*, *287*(7), 4740–4751.

Geiss-Friedlander, R., & Melchior, F. (2007). Concepts in sumoylation: a decade on. *Nature Reviews Molecular Cell Biology*, *8*(12), 947–956.

George Priya Doss, C., Rajith, B., Chakraboty, C., Balaji, V., Magesh, R., Gowthami, B., … Das, M. (2014). In silico profiling and structural insights of missense mutations in RET protein kinase domain by molecular dynamics and docking approach. *Molecular bioSystems*, *10*(3), 421–436.

Gilis, D., & Rooman, M. (2000). PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. *Protein Engineering*, *13*(12), 849–856.

Giver, L., Gershenson, A., Freskgard, P. O., & Arnold, F. H. (1998). Directed evolution of a thermostable esterase. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(22), 12809–12813.

Gnad, F., Baucom, A., Mukhyala, K., Manning, G., & Zhang, Z. (2013). Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics*, *14 Suppl 3*, S7. https://doi.org/10.1186/1471-2164-14-S3-S7

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabási, A.-L. (2007).

The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(21), 8685–90. https://doi.org/10.1073/pnas.0701361104

Goldenzweig, A., Goldsmith, M., Hill, S. E., Gertman, O., Laurino, P., Ashani, Y., … Fleishman, S. J. (2016). Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Molecular Cell*, *63*(2), 337–346.

Gonzalez-Perez, A., Deu-Pons, J., & Lopez-Bigas, N. (2012). Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Medicine*, *4*(11), 89.

Gonzalez-Perez, A., & Lopez-Bigas, N. (2012). Functional impact bias reveals cancer drivers. *Nucleic Acids Research*, *40*(21), e169.

Gonz âlez-P érez, A., & L ópez-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *American Journal of Human Genetics*, *88*(4), 440–9. https://doi.org/10.1016/j.ajhg.2011.03.004

Gonzalez-Perez, A., Mustonen, V., Reva, B., Ritchie, G. R. S., Creixell, P., Karchin, R., … International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group. (2013). Computational approaches to identify functional genetic variants in cancer genomes. *Nature Methods*, *10*(8), 723–729.

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., … Lopez-Bigas, N. (2013). IntOGen-mutations identifies

cancer drivers across tumor types. *Nature Methods*, *10*(11), 1081–1082.

Greenfield, N., & Fasman, G. D. (1969). Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry*, *8*(10), 4108–4116.

Grigoryan, G., Reinke, A. W., & Keating, A. E. (2009). Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature*, *458*(7240), 859–864.

Gryfe, R., & Gallinger, S. (2001). Microsatellite instability, mismatch repair deficiency, and colorectal cancer. *Surgery*, *130*(1), 17–20.

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, … Montgomery. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204–213.

Guerois, R., Nielsen, J. E., & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of Molecular Biology*, *320*(2), 369–387.

Gusev, A., Lee, S. H., Trynka, G., Finucane, H., Vilhjálmsson, B. J., Xu, H., … SWE-SCZ Consortium. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *American Journal of Human Genetics*, *95*(5), 535–552.

Hecht, M., Bromberg, Y., & Rost, B. (2015). Better prediction of functional effects for sequence variants. *BMC Genomics*, *16*(Suppl 8), S1. https://doi.org/10.1186/1471-2164-16-S8-S1

Heselpoth, R. D., Yin, Y., Moult, J., & Nelson, D. C. (2015). Increasing the stability

of the bacteriophage endolysin PlyC using rationale-based FoldX computational modeling. *Protein Engineering, Design & Selection: PEDS*, *28*(4), 85–92.

Hofree, M., Carter, H., Kreisberg, J. F., Bandyopadhyay, S., Mischel, P. S., Friend, S., & Ideker, T. (2016). Challenges in identifying cancer genes by analysis of exome sequencing data. *Nature Communications*, *7*, 12096.

Horn, S., Figl, A., Rachakonda, P. S., Fischer, C., Sucker, A., Gast, A., … Kumar, R. (2013). TERT promoter mutations in familial and sporadic melanoma. *Science*, *339*(6122), 959–961.

Hoskins, R. A., Repo, S., Barsky, D., Andreoletti, G., Moult, J., & Brenner, S. E. (2017). Reports from CAGI: The Critical Assessment of Genome Interpretation. *Human Mutation*, *38*(9), 1039–1041.

Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V., Chin, L., & Garraway, L. A. (2013). Highly recurrent TERT promoter mutations in human melanoma. *Science*, *339*(6122), 957–959.

Hwang, D. G., & Green, P. (2004). Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(39), 13994–14001.

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., … Sieh, W. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *The American Journal of Human Genetics*, *99*(4), 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Röthlisberger, D., Zanghellini,

A., … Baker, D. (2008). De novo computational design of retro-aldol enzymes. *Science*, *319*(5868), 1387–1391.

Jones, D. T., & Cozzetto, D. (2015). DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*, *31*(6), 857–863.

Jonsson, P. F., & Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics*, *22*(18), 2291–2297.

Kamal, M. Z., Mohammad, T. A., Krishnamoorthy, G., & Rao, N. M. (2012). Role of active site rigidity in activity: MD simulation and fluorescence study on a lipase mutant. *PLoS One*, *7*(4), e35188.

Kaminker, J. S., Zhang, Y., Watanabe, C., & Zhang, Z. (2007). CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Research*, *35*(Web Server issue), W595-598.

Kaminker, J. S., Zhang, Y., Waugh, A., Haverty, P. M., Peters, B., Sebisanovic, D., … Zhang, Z. (2007). Distinguishing cancer-associated missense mutations from common polymorphisms. *Cancer Research*, *67*(2), 465–473.

Kanaya, S., Oobatake, M., & Liu, Y. (1996). Thermal stability of Escherichia coli ribonuclease HI and its active site mutants in the presence and absence of the Mg2+ ion. Proposal of a novel catalytic role for Glu48. *The Journal of Biological Chemistry*, *271*(51), 32729–32736.

Kar, G., Gursoy, A., & Keskin, O. (2009). Human cancer protein-protein interaction network: A structural perspective. *PLoS Computational Biology*, *5*(12), e1000601.

Katsonis, P., & Lichtarge, O. (2014). A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. *Genome Research*, *24*(12), 2050–8. https://doi.org/10.1101/gr.176214.114

Kellogg, E. H., Leaver-Fay, A., & Baker, D. (2011). Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins*, *79*(3), 830–838.

Kelly, S. M., Jess, T. J., & Price, N. C. (2005). How to study proteins by circular dichroism. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1751*(2), 119–139.

Khan, S., & Vihinen, M. (2010). Performance of protein stability predictors. *Human Mutation*, *31*(6), 675–684.

Khurana, E., Fu, Y., Colonna, V., Mu, X. J., Kang, H. M., Lappalainen, T., … Gerstein, M. (2013). Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science*, *342*(6154), 1235587.

Kircher, M., Witten, D., Jain, P., O'Roak, B., Cooper, G., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315.

Kloosterman, W. P., Francioli, L. C., Hormozdiari, F., Marschall, T., Hehir-Kwa, J. Y., Abdellaoui, A., … Guryev, V. (2015). Characteristics of de novo structural changes in the human genome. *Genome Research*, *25*(6), 792–801.

Kondrashov, A. S., Sunyaev, S., & Kondrashov, F. A. (2002). Dobzhansky-Muller incompatibilities in protein evolution. *Proceedings of the National Academy of*

Sciences of the United States of America, *99*(23), 14878–14883.

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., … Stefansson, K. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, *488*(7412), 471–475.

Krause, R. M. (1957). Studies on bacteriophages of hemolytic streptococci. I. Factors influencing the interaction of phage and susceptible host cell. *The Journal of Experimental Medicine*, *106*(3), 365–384.

Kryukov, G. V., Pennacchio, L. A., & Sunyaev, S. R. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *The American Journal of Human Genetics*, *80*(4), 727–739.

Kuhlman, B., & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(19), 10383–10388.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, *302*(5649), 1364–1368.

Kumar, R. D., Searleman, A. C., Swamidass, S. J., Griffith, O. L., & Bose, R. (2015). Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data. *Bioinformatics*, *31*(22), 3561–3568.

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, *227*(5259), 680–685.

Lam, S. Y., Yeung, R. C., Yu, T. H., Sze, K. H., & Wong, K. B. (2011). A rigidifying salt-bridge favors the activity of thermophilic enzyme at high temperatures at the

expense of low-temperature activity. *PLoS Biology*, *9*(3), e1001027.

Lamandé, S. R., Bateman, J. F., Hutchison, W., McKinlay Gardner, R. J., Bower, S. P., Byrne, E., & Dahl, H. H. (1998). Reduced collagen VI causes Bethlem myopathy: a heterozygous COL6A1 nonsense mutation results in mRNA decay and functional haploinsufficiency. *Human Molecular Genetics*, *7*(6), 981–989.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., … Maglott, D. R. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, *44*(D1), D862-868.

Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., … Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, *505*(7484), 495–501.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V, Cibulskis, K., Sivachenko, A., … Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, *499*(7457), 214–218.

Lee-Chen, G. J., Lin, S. P., Lin, S. Z., Chuang, C. K., Hsiao, K. T., Huang, C. F., & Lien, W. C. (2002). Identification and characterisation of mutations underlying Sanfilippo syndrome type B (mucopolysaccharidosis type IIIB). *Journal of Medical Genetics*, *39*(2), E3. https://doi.org/10.1136/JMG.39.2.E3

Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V, Thomas, J. L., … Raphael, B. J. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, *47*(2), 106–114.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., …

Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291.

Li, B., Krishnan, V. G., Mort, M. E., Xin, F., Kamati, K. K., Cooper, D. N., … Radivojac, P. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics (Oxford, England)*, *25*(21), 2744–50. https://doi.org/10.1093/bioinformatics/btp528

Li, X., Kim, Y., Tsang, E. K., Davis, J. R., Damani, F. N., Chiang, C., … Montgomery, S. B. (2017). The impact of rare variation on gene expression across tissues. *Nature*, *550*(7675), 239–243.

Lichtarge, O., Bourne, H. R., & Cohen, F. E. (1996). An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families. *Journal of Molecular Biology*, *257*(2), 342–358.

Lippow, S. M., Wittrup, K. D., & Tidor, B. (2007). Computational design of antibody-affinity improvement beyond in vivo maturation. *Nature Biotechnology*, *25*(10), 1171–1176.

Liu, J., Faeder, J. R., & Camacho, C. J. (2009). Toward a quantitative theory of intrinsically disordered proteins and their function. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(47), 19819–19823.

Liu, X., Jian, X., & Boerwinkle, E. (2013). dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Human Mutation*, *34*(9), E2393-2402.

Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences of the United States of*

America, *107*(3), 961–968.

Manfredi, J. J. (2010). The Mdm2-p53 relationship evolves: Mdm2 swings both ways as an oncogene and a tumor suppressor. *Genes & Development*, *24*(15), 1580–1589.

Mao, Y., Chen, H., Liang, H., Meric-Bernstam, F., Mills, G. B., & Chen, K. (2013). CanDrA: cancer-specific driver missense mutation annotation with optimized features. *PLoS One*, *8*(10), e77945.

Martelotto, L. G., Ng, C. K., De Filippo, M. R., Zhang, Y., Piscuoglio, S., Lim, R. S., … Weigelt, B. (2014). Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. *Genome Biology*, *15*(10), 484.

Martincorena, I., & Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science*, *349*(6255), 1483–1489.

Martincorena, I., & Luscombe, N. M. (2013). Non-random mutation: The evolution of targeted hypermutation and hypomutation. *BioEssays*, *35*(2), 123–130.

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., … Campbell, P. J. (2015). High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, *348*(6237), 880–886.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., … Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, *337*(6099), 1190–1195.

McFarland, C. D., Korolev, K. S., Kryukov, G. V, Sunyaev, S. R., & Mirny, L. A. (2013). Impact of deleterious passenger mutations on cancer progression.

Proceedings of the National Academy of Sciences of the United States of America, *110*(8), 2910–5. https://doi.org/10.1073/pnas.1213968110

McGowan, S., Buckle, A. M., Mitchell, M. S., Hoopes, J. T., Gallagher, D. T., Heselpoth, R. D., … Nelson, D. C. (2012). X-ray crystal structure of the streptococcal specific phage lysin PlyC. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(31), 12752–12757.

Meiering, E. M., Serrano, L., & Fersht, A. R. (1992). Effect of active site residues in barnase on activity and stability. *Journal of Molecular Biology*, *225*(3), 585–589.

Meiyappan, M., Concino, M, F., & Norton, A, W. (2014). *US08775146B2*.

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, *12*(4), R41.

Michaelson, J. J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., … Sebat, J. (2012). Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell*, *151*(7), 1431–1442.

Midic, U., Oldfield, C. J., Dunker, A. K., Obradovic, Z., & Uversky, V. N. (2009). Protein disorder in the human diseasome: unfoldomics of human genetic diseases. *BMC Genomics*, *10 Suppl 1*(1), S12.

Molodecky, N. A., Soon, I. S., Rabi, D. M., Ghali, W. A., Ferris, M., Chernoff, G., … Kaplan, G. G. (2012). Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology*, *142*(1),

46–54.e42.

Moult, J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology*, *15*(3), 285–289.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins, 84 Suppl 1*, 4–14.

Mukaiyama, A., Haruki, M., Ota, M., Koga, Y., Takano, K., & Kanaya, S. (2006). A hyperthermophilic protein acquires function at the cost of stability. *Biochemistry*, *45*(42), 12673–12679.

NCBI Resource Coordinators. (2015). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *44*(D1), D7–D19. https://doi.org/10.1093/nar/gkv1290

Nelson, D., Loomis, L., & Fischetti, V. A. (2001). Prevention and elimination of upper respiratory colonization of mice by group A streptococci by using a bacteriophage lytic enzyme. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(7), 4107–4112.

Nelson, D., Schuch, R., Chahales, P., Zhu, S., & Fischetti, V. A. (2006). PlyC: A multimeric bacteriophage lysin. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(28), 10765–10770.

Nelson, D., Schuch, R., Zhu, S., Tscherne, D. M., & Fischetti, V. A. (2003). Genomic sequence of C1, the first streptococcal phage. *Journal of Bacteriology*, *185*(11), 3325–3332.

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect

protein function. *Nucleic Acids Research*, *31*(13), 3812–3814.

Ng, P. C., Levy, S., Huang, J., Stockwell, T. B., Walenz, B. P., Li, K., … Venter, J. C. (2008). Genetic variation in an individual human exome. *PLoS Genetics*, *4*(8), e1000160.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., & Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genetics*, *6*(4), e1000888.

Niroula, A., Urolagin, S., & Vihinen, M. (2015). PON-P2: Prediction Method for Fast and Reliable Identification of Harmful Variants. *PLOS ONE*, *10*(2), e0117380. https://doi.org/10.1371/journal.pone.0117380

Niroula, A., & Vihinen, M. (2016). Variation Interpretation Predictors: Principles, Types, Performance, and Choice. *Human Mutation*, *37*(6), 579–597.

Nishi, H., Tyagi, M., Teng, S., Shoemaker, B. A., Hashimoto, K., Alexov, E., … Panchenko, A. R. (2013). Cancer missense mutations alter binding properties of proteins and their interaction networks. *PLoS One*, *8*(6), e66273.

O'Brien, J. S. (1972). Sanfilippo syndrome: profound deficiency of alpha-acetylglucosaminidase activity in organs and skin fibroblasts from type-B patients. *Proceedings of the National Academy of Sciences of the United States of America*, *69*(7), 1720–2.

Ohage, E., & Steipe, B. (1999). Intrabody construction and expression. I. The critical role of VL domain stability. *Journal of Molecular Biology*, *291*(5), 1119–1128.

Olatubosun, A., Väliaho, J., Härkönen, J., Thusberg, J., & Vihinen, M. (2012). PON-P: Integrated predictor for pathogenicity of missense variants. *Human Mutation*,

33(8), 1166–1174. https://doi.org/10.1002/humu.22102

Oldfield, C. J., Meng, J., Yang, J. Y., Yang, M. Q., Uversky, V. N., & Dunker, A. K. (2008). Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics*, *9*(Suppl 1), S1.

Ollikainen, N., Smith, C. A., Fraser, J. S., & Kortemme, T. (2013). Flexible backbone sampling methods to model and design protein alternative conformations. *Methods in Enzymology*, *523*, 61–85.

Orr, H. A. (2009). Fitness and its role in evolutionary genetics. *Nature Reviews Genetics*, *10*(8), 531–539. https://doi.org/10.1038/nrg2603

Pajkos, M., Mészáros, B., Simon, I., & Dosztányi, Z. (2012). Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. *Molecular bioSystems*, *8*(1), 296–307.

Pal, L. R., & Moult, J. (2015). Genetic Basis of Common Human Disease: Insight into the Role of Missense SNPs from Genome-Wide Association Studies. *Journal of Molecular Biology*, *427*(13), 2271–2289.

Palles, C., Cazier, J.-B., Howarth, K. M., Domingo, E., Jones, A. M., Broderick, P., … Tomlinson, I. (2013). Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature Genetics*, *45*(2), 136–144.

Pantoliano, M. W., Whitlow, M., Wood, J. F., Dodd, S. W., Hardman, K. D., Rollence, M. L., & Bryan, P. N. (1989). Large increases in general stability for subtilisin BPN' through incremental changes in the free energy of unfolding. *Biochemistry*, *28*(18), 7205–7213.

Parthiban, V., Gromiha, M. M., Hoppe, C., & Schomburg, D. (2007). Structural analysis and prediction of protein mutant stability using distance and torsion potentials: role of secondary structure and solvent accessibility. *Proteins*, *66*(1), 41–52.

Parthiban, V., Gromiha, M. M., & Schomburg, D. (2006). CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Research*, *34*(Web Server issue), W239-242.

Peterson, T. A., Doughty, E., & Kann, M. G. (2013). Towards precision medicine: advances in computational approaches for the analysis of human variants. *Journal of Molecular Biology*, *425*(21), 4047–4063.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, *20*(1), 110–121.

Potapov, V., Cohen, M., & Schreiber, G. (2009). Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Engineering, Design & Selection : PEDS*, *22*(9), 553–560.

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics*, *69*(1), 124–137.

Procko, E., Hedman, R., Hamilton, K., Seetharaman, J., Fleishman, S. J., Su, M., … Baker, D. (2013). Computational design of a protein-based enzyme inhibitor. *Journal of Molecular Biology*, *425*(18), 3563–3575.

Provencher, S. W., & Glockner, J. (1981). Estimation of globular protein secondary structure from circular dichroism. *Biochemistry*, *20*(1), 33–37.

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., … Baker, D. (2009). Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins, 77 Suppl 9*, 89–99.

Redler, R. L., Das, J., Diaz, J. R., & Dokholyan, N. V. (2016). Protein Destabilization as a Common Factor in Diverse Inherited Disorders. *Journal of Molecular Evolution*, *82*(1), 11–16.

Reimand, J., & Bader, G. D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Molecular Systems Biology*, *9*(1), 637.

Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-Throughput Sequencing Technologies. *Molecular Cell*, *58*(4), 586–597.

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

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., … ACMG Laboratory Quality Assurance Committee. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, *17*(5), 405–423.

Roach, J. C., Glusman, G., Smit, A. F. A., Huff, C. D., Hubley, R., Shannon, P. T., … Galas, D. J. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, *328*(5978), 636–639.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., … Grant,

G. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., & Baker, D. (2004). Protein structure prediction using Rosetta. *Methods in Enzymology*, *383*, 66–93.

Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D., & Bolon, D. N. (2013). Analyses of the Effects of All Ubiquitin Point Mutants on Yeast Growth Rate. *Journal of Molecular Biology*, *425*(8), 1363–1377. https://doi.org/10.1016/j.jmb.2013.01.032

Rost, B., & Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins: Structure, Function, and Genetics*, *20*(3), 216–226. https://doi.org/10.1002/prot.340200303

Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., … Baker, D. (2008). Kemp elimination catalysts by computational enzyme design. *Nature*, *453*(7192), 190–195.

Rubin, A. F., Gelman, H., Lucas, N., Bajjalieh, S. M., Papenfuss, A. T., Speed, T. P., & Fowler, D. M. (2017). A statistical framework for analyzing deep mutational scanning data. *Genome Biology*, *18*(1), 150.

Rubio-Perez, C., Tamborero, D., Schroeder, M. P., Antol ń, A. A., Deu-Pons, J., Perez-Llamas, C., … Lopez-Bigas, N. (2015). In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell*, *27*(3), 382–396.

Sabarinathan, R., Pich, O., Martincorena, I., Rubio-Perez, C., Juul, M., Wala, J., …

Net, I. P.-C. A. of W. G. (2017). The whole-genome panorama of cancer drivers. *bioRxiv*, 190330.

Samish, I. (2009). *Search and sampling in structural bioinformatics.* (J. Gu & P. E. Bourne, Eds.), *Structural Bioinformatics*. New York: Wiley-Blackwell.

Samish, I. (Ed.). (2017). *Computational Protein Design* (Vol. 1529). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-6637-0

Sanfilippo, S. J., Podosin, R., Langer, L., & Good, R. A. (1963). Mental retardation associated with acid mucopolysaccharidurie (heparitin sulfate type). *The Journal of Pediatrics*, *63*(4), 837–838. https://doi.org/10.1016/S0022-3476(63)80279-6

Sanz, J. M., Garcia, J. L., Laynez, J., Usobiaga, P., & Menendez, M. (1993). Thermal stability and cooperative domains of CPL1 lysozyme and its NH2- and COOH-terminal modules. Dependence on choline binding. *The Journal of Biological Chemistry*, *268*(9), 6125–6130.

Schuster-Böckler, B., & Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, *488*(7412), 504–507.

Schwarz, J. M., Rödelsperger, C., Schuelke, M., & Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods*, *7*(8), 575–576. https://doi.org/10.1038/nmeth0810-575

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Research*, *33*(Web Server issue), W382-388.

Schymkowitz, J., Rousseau, F., Martins, I. C., Ferkinghoff-Borg, J., Stricher, F., &

Serrano, L. (2005). Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(29), 10147–10152.

Seeliger, D., & de Groot, B. L. (2010). Protein thermostability calculations using alchemical free energy simulations. *Biophysical Journal*, *98*(10), 2309–2316.

Ségurel, L., Wyman, M. J., & Przeworski, M. (2014). Determinants of mutation rate variation in the human germline. *Annual Review of Genomics and Human Genetics*, *15*(1), 47–70.

Serrano, L., Day, A. G., & Fersht, A. R. (1993). Step-wise mutation of barnase to binase. A procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability. *Journal of Molecular Biology*, *233*(2), 305–312.

Shah, P. S., Hom, G. K., Ross, S. A., Lassila, J. K., Crowhurst, K. A., & Mayo, S. L. (2007). Full-sequence computational design and solution structure of a thermostable protein variant. *Journal of Molecular Biology*, *372*(1), 1–6.

Shaw, J. E., Sicree, R. A., & Zimmet, P. Z. (2010). Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Research and Clinical Practice*, *87*(1), 4–14.

Shendure, J., & Akey, J. M. (2015). The origins, determinants, and consequences of human mutations. *Science*, *349*(6255), 1478–1483.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145.

Sherman, S. L., Petersen, M. B., Freeman, S. B., Hersey, J., Pettay, D., Taft, L., …

Hassold, T. J. (1994). Non-disjunction of chromosome 21 in maternal meiosis I: evidence for a maternal age-dependent mechanism involving reduced recombination. *Human Molecular Genetics*, *3*(9), 1529–1535.

Shi, Z., & Moult, J. (2011). Structural and Functional Impact of Cancer-Related Missense Somatic Mutations. *Journal of Molecular Biology*, *413*(2), 495–512.

Shi, Z., Sellers, J., & Moult, J. (2012). Protein stability and in vivo concentration of missense mutations in phenylalanine hydroxylase. *Proteins: Structure, Function, and Bioinformatics*, *80*(1), 61–70.

Shih, P., & Kirsch, J. F. (1995). Design and structural analysis of an engineered thermostable chicken lysozyme. *Protein Science*, *4*(10), 2063–2072.

Shihab, H. A., Gough, J., Cooper, D. N., Day, I. N. M., & Gaunt, T. R. (2013). Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics*, *29*(12), 1504–1510.

Shoichet, B. K., Baase, W. A., Kuroki, R., & Matthews, B. W. (1995). A relationship between protein stability and protein function. *Proceedings of the National Academy of Sciences*, *92*(2), 452–456.

Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St Clair, J. L., … Baker, D. (2010). Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science*, *329*(5989), 309–313.

Soon, W. W., Hariharan, M., & Snyder, M. P. (2013). High-throughput sequencing for biology and medicine. *Molecular Systems Biology*, *9*(1), 640.

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S. T., … Cooper, D. N. (2003). Human Gene Mutation Database (HGMD): 2003

update. *Human Mutation*, *21*(6), 577–581.

Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., … Cooper, D. N. (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics*, *136*(6), 665–677.

Stenson, P. D., Mort, M., Ball, E. V., Shaw, K., Phillips, A. D., & Cooper, D. N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics*, *133*(1), 1–9.

Stenson, P. D., Mort, M., Ball, E. V, Howells, K., Phillips, A. D., Thomas, N. S., & Cooper, D. N. (2009). The Human Gene Mutation Database: 2008 update. *Genome Medicine*, *1*(1), 13.

Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., … Campbell, P. J. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, *144*(1), 27–40.

Strauch, E.-M., Fleishman, S. J., & Baker, D. (2014). Computational design of a pH-sensitive IgG binding protein. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(2), 675–680.

Tamborero, D., Gonzalez-Perez, A., & Lopez-Bigas, N. (2013). OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, *29*(18), 2238–2244.

Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C.,

Reimand, J., … Lopez-Bigas, N. (2013). Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific Reports*, *3*(1), 2650.

Tessitore, A., Villani, G. R., Di Domenico, C., Filocamo, M., Gatti, R., & Di Natale, P. (2000). Molecular defects in the α-N-acetylglucosaminidase gene in Italian Sanfilippo type B patients. *Human Genetics*, *107*(6), 568–576. https://doi.org/10.1007/s004390000429

Thomas, P. D., Kejariwal, A., Guo, N., Mi, H., Campbell, M. J., Muruganujan, A., & Lazareva-Ulitsky, B. (2006). Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Research*, *34*(Web Server), W645–W650. https://doi.org/10.1093/nar/gkl229

Thusberg, J., Olatubosun, A., & Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Human Mutation*, *32*(4), 358–368.

Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., & Karchin, R. (2016). Evaluating the evaluation of cancer driver genes. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(50), 14330–14335.

Tomasetti, C., Li, L., & Vogelstein, B. (2017). Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, *355*(6331), 1330–1334.

UniProt Consortium, T. U. (2015). UniProt: a hub for protein information. *Nucleic*

Acids Research, *43*(Database issue), D204-12. https://doi.org/10.1093/nar/gku989

Vacic, V., Markwick, P. R. L., Oldfield, C. J., Zhao, X., Haynes, C., Uversky, V. N., & Iakoucheva, L. M. (2012). Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. *PLoS Computational Biology*, *8*(10), e1002709.

Valstar, M. J., Ruijter, G. J. G., van Diggelen, O. P., Poorthuis, B. J., & Wijburg, F. A. (2008). Sanfilippo syndrome: A mini-review. *Journal of Inherited Metabolic Disease*, *31*(2), 240–252.

Varea, J., Monterroso, B., Sáiz, J. L., López-Zumel, C., García, J. L., Laynez, J., … Menéndez, M. (2004). Structural and thermodynamic characterization of Pal, a phage natural chimeric lysin active against pneumococci. *The Journal of Biological Chemistry*, *279*(42), 43697–43707.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. *Science*, *339*(6127), 1546–1558.

von der Osten, C., Branner, S., Hastrup, S., Hedegaard, L., Rasmussen, M. D., Bisgard-Frantzen, H., … Mikkelsen, J. M. (1993). Protein engineering of subtilisins to improve stability in detergent formulations. *Journal of Biotechnology*, *28*(1), 55–68.

von Figura, K., & Kresse, H. (1972). The Sanfilippo B corrective factor: A N-acetyl-α-D-glucosaminidase. *Biochemical and Biophysical Research Communications*, *48*(2), 262–269. https://doi.org/10.1016/S0006-291X(72)80044-5

Wang, I. N., Smith, D. L., & Young, R. (2000). Holins: the protein clocks of bacteriophage infections. *Annual Review of Microbiology*, *54*, 799–825.

Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Human Mutation*, *17*(4), 263–270.

Ward, A. J., & Cooper, T. A. (2009). The pathobiology of splicing. *The Journal of Pathology*, *220*(2), 152–163.

Weber, B., Guo, X.-H., Kleijer, W. J., van de Kamp, J. J., Poorthuis, B. J., & Hopwood, J. J. (1999). Sanfilippo type B syndrome (mucopolysaccharidosis III B): allelic heterogeneity corresponds to the wide spectrum of clinical phenotypes. *European Journal of Human Genetics*, *7*(1), 34–44. https://doi.org/10.1038/sj.ejhg.5200242

Weile, J., Sun, S., Cote, A. G., Knapp, J., Verby, M., Mellor, J. C., … Roth, F. P. (2017). Expanding the Atlas of Functional Missense Variation for Human Genes. *Doi.org*, 166595. https://doi.org/10.1101/166595

Whitmore, L., & Wallace, B. A. (2004). DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic Acids Research*, *32*(Web Server issue), W668-673.

Wickham H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag. https://doi.org/10.1007/978-0-387-98141-3

Wong, W. C., Kim, D., Carter, H., Diekhans, M., Ryan, M. C., & Karchin, R. (2011). CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics*, *27*(15), 2147–2148.

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjöblom, T., Leary, R. J., …

Vogelstein, B. (2007). The Genomic Landscapes of Human Breast and Colorectal Cancers. *Science*, *318*(5853), 1108–1113.

Wrenbeck, E. E., Faber, M. S., & Whitehead, T. A. (2017). Deep sequencing methods for protein engineering and design. *Current Opinion in Structural Biology*, *45*, 36–44.

Xie, Y., An, J., Yang, G., Wu, G., Zhang, Y., Cui, L., & Feng, Y. (2014). Enhanced enzyme kinetic stability by increasing rigidity within the active site. *The Journal of Biological Chemistry*, *289*(11), 7994–8006.

Xiong, P., Wang, M., Zhou, X., Zhang, T., Zhang, J., Chen, Q., & Liu, H. (2014). Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. *Nature Communications*, *5*, 5330.

Xue, Y., Chen, Y., Ayub, Q., Huang, N., Ball, E. V., Mort, M., … 1000 Genomes Project Consortium. (2012). Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *The American Journal of Human Genetics*, *91*(6), 1022–1032.

Yin, Y., Kundu, K., Pal, L. R., & Moult, J. (2017). Ensemble variant interpretation methods to predict enzyme activity and assign pathogenicity in the CAGI4 NAGLU (Human N-acetyl-glucosaminidase) and UBE2I (Human SUMO-ligase) challenges. *Human Mutation*, *38*(9), 1109–1122.

Young, R. (1992). Bacteriophage lysis: mechanism and regulation. *Microbiological Reviews*, *56*(3), 430–481.

Yue, P., Forrest, W. F., Kaminker, J. S., Lohr, S., Zhang, Z., & Cavet, G. (2010).

Inferring the functional effects of mutation through clusters of mutations in homologous proteins. *Human Mutation*, *31*(3), 264–71. https://doi.org/10.1002/humu.21194

Yue, P., Li, Z., & Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of Molecular Biology*, *353*(2), 459–473.

Yue, P., & Moult, J. (2006). Identification and analysis of deleterious human SNPs. *Journal of Molecular Biology*, *356*(5), 1263–1274.

Yutani, K., Ogasahara, K., Tsujita, T., & Sugino, Y. (1987). Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit. *Proceedings of the National Academy of Sciences of the United States of America*, *84*(13), 4441–4444.

Zhang, J., Kinch, L. N., Cong, Q., Weile, J., Sun, S., Cote, A. G., … Grishin, N. V. (2017). Assessing predictions of fitness effects of missense mutations in SUMO-conjugating enzyme UBE2I. *Human Mutation*, *38*(9), 1051–1063. https://doi.org/10.1002/humu.23293

Zhi, W., Srere, P. A., & Evans, C. T. (1991). Conformational stability of pig citrate synthase and some active-site mutants. *Biochemistry*, *30*(38), 9281–9286.

Zhou, H., & Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science*, *11*(11), 2714–2726.