

## ABSTRACT

Title of Document: COMPUTATIONAL METHODS IN PROTEIN STRUCTURE, EVOLUTION AND NETWORKS.

Chen Cao, Doctor of Philosophy, 2013

Directed By: Professor John Moult, Institute for Bioscience and Biotechnology Research. Department of Cell Biology and Molecular Genetics, University of Maryland

The advent of new sequencing technology has resulted in the accumulation of a large amount of information on human DNA variation. In order to make sense of these data in the context of biology and medicine, new methods are needed both for analysis and for integration with other resources. In this work: 1) I studied the distribution pattern of human DNA variants across populations using data from the 1000 genomes project and investigated several evolutionary biology questions from the perspective of population genomics. I found population level support for trends previously observed between species, including selection against deleterious variants, and lower frequency of variants in highly expressed genes and highly connected genes. I was also able to show that the correlation between synonymous and non-synonymous variant levels is

a consequence of both mutation prevalence variation across the genome and shared selection pressure. 2) I performed a systematic evaluation of the effectiveness of GWAS (Genome Wide Association Studies) for finding potential drug targets and discovered the method is very ineffective for this purpose. I proposed two reasons to explain this finding, selection against variants in drug targets and the relatively short length of drug target genes. I discovered that GWAS genes and drug targets are closely associated in the biological network, and on that basis, developed a machine learning algorithm to leverage the GWAS results for the identification of potential drug targets, making use of biological network information. As a result, I identified some potential drug repurposing opportunities. 3) I developed a method to increase the number of protein structure models available for interpreting the impact of human non-synonymous variants, important for not only the understanding the mechanisms of genetic disease but also in the study of human protein evolution. The method enables the impact of approximately 40% more missense variants to be reliably modeled. In summary, these three projects demonstrate that value of computational methods in addressing a wide range of problems in protein structure, evolution, and networks.

COMPUTATIONAL METHODS IN PROTEIN STRUCTURE, EVOLUTION AND  
NETWORKS

By

Chen Cao

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  
2013

Advisory Committee:  
Professor John Moult, Chair  
Professor Philip Bryan  
Professor Lindley Darden, Dean's representative  
Associated Professor Eric Haag  
Associated Professor Mihai Pop

© Copyright by  
Chen Cao  
2013

# Dedication

To my parents

## Acknowledgements

First of all, I would like to give my deepest gratitude to my Ph.D. advisor Professor John Moulton. He guided me through the fascinating world of computational biology by endless inspirations and support. During these years of study, he always encouraged me to explore new ideas, and when I got lost in the research, he was the one to get me back on track. To me, he is both an outstanding mentor and a great friend. Being his student is one of the luckiest and most honorable things in my life.

I would also give my unreserved appreciation to my committee members: Professor Lindley Darden, Professor Philip Bryan, Professor Eric Haag, and Professor Mihai Pop. They are excellent researchers in their own fields and they inspired me with great ideas and suggestions from their unique perspectives, which is very valuable to my research.

I also want to thank Professor Arlin Stolfus for the inspiring discussions about evolutionary biology and Dr. Peter Stein for his generous help in high performance computing. I want to thank Dr. Lipika Ray and Dr. Zhen Shi for giving me comments and advice as senior members in the lab and I also want to thank Mr. Yizhou Yin and Mr. Albert Yu, who are both my lab-mates and roommates, for their tremendous help in my study and life.

I have to express my gratitude to my Mother. Although she is far from here, I can

always feel her priceless support and I know her heart is always with me. My special thanks to my father, his bravery and optimism in fighting against disease always motivate me to face difficulties in my life and study with confidence and courage. Last but not least, I'd like to thank my girlfriend Yulan Xia – this long and arduous journey became such a wonderful experience with the companionship of you!

# Table of Contents

Dedication.....	ii
Acknowledgements.....	iii
Table of Contents.....	v
Chapter 1: Introduction.....	1
1.1: Human single nucleotide variants and evolutionary biology.....	1
1.2: GWAS and drug targets.....	4
1.3: Protein structure modeling.....	6
1.4: Overview.....	8
Chapter 2: Distribution of human single nucleotide variants and implications for protein sequence evolution.....	9
2.1: Introduction.....	9
2.2: Results.....	17
Deleterious variants occur at lower frequencies than neutral variants. ....	17
Correlation between non-synonymous and synonymous variant density.....	18
ns-SNV densities are restrained by the mRNA expression level.....	24
ns-SNV densities also correlate with protein network features.....	27
Outlier genes.....	30
2.3: Methods.....	34
1000 genomes project data.....	34

Human mRNA expression data .....	35
SNV impact prediction .....	35
Human Protein Functional Interaction Network.....	36
Partial correlation.....	36
Quantitative estimation of the selection pressure caused by mRNA expression level and number of functional interactions.....	36
GO enrichment analysis of outliers.....	37
2.4: Discussion.....	37
Chapter 3: GWAS and drug targets .....	42
3.1: Introduction.....	42
3.2: Results.....	44
Comparison of the GWAS catalog and Drugbank shows GWAS only detects a very small fraction of existing drug targets. ....	44
Analysis using 1000 genomes project data shows Drug Target genes have fewer high frequency deleterious non-synonymous SNPs than GWAS reported genes .....	48
Evolutionary analysis shows drug target genes are under slightly stronger negative selection than GWAS reported genes.....	50
The influence of transcript length.....	52
Network analysis shows GWAS reported genes are close to drug target genes in a biological network.....	53
Machine learning method for drug target discovery.....	60
Potential new drug targets for drug repurposing.....	64

3.3: Methods .....	68
Connecting GWAS reported genes with drug targets through the drug indication information from Drugbank.....	68
Calculating expected overlap between GWAS reported genes and drug targets using a complete random model .....	69
SNP impact analysis for GWAS genes and drug target genes.....	69
Transcript length analysis .....	70
Evolutionary analysis for GWAS reported genes and Drug target genes.....	71
Human gene network analysis for GWAS reported genes and drug target genes .....	71
Machine learning for drug targets.....	72
3.4: Discussion.....	72
Chapter 4: Improved missense variant modeling accuracy using an Alignment Quality Estimator.....	76
4.1: Introduction.....	76
4.2: Results.....	79
Score Drop is a good criterion for Alignment quality .....	79
Comparison between profile-profile alignment and substitution matrix based alignment.....	82
Probability of a correct alignment as a function of score drop .....	83
SNPs3D performs better at high confidence alignment positions than at the low confidence ones.....	87
Examples of low sequence identity, high quality models.....	89

Increasing the number of model-able variants by picking high confidence alignments in low sequence identity models .....	93
4.3: Methods .....	93
Extracting pair-wise alignments from Balibase .....	93
Comparing Smith-Waterman alignments with extracted pair-wise alignments .	94
Force-Bypass algorithm .....	95
Annotating missense mutations in the 1000 genome data .....	98
Structural modeling of missense variants .....	98
Measuring the performance of variant impact modeling for different sets of models .....	99
4.4: Discussion .....	101
Chapter 5: Conclusions and perspectives .....	104
5.1: Population Genomics: studying evolution in detail .....	104
5.2: Integration of GWAS with other knowledge and the future of GWAS with rare variants .....	106
5.3 Protein modeling and the future of personal medicine .....	108
Bibliography .....	110

## Chapter 1: Introduction

In recent years, Next Generation Sequencing (Bentley, Balasubramanian et al. 2008; Eid, Fehr et al. 2009; Rothberg, Hinz et al. 2011) has greatly accelerated the generation of biological sequencing data. The explosive amount of sequencing data, combined with various other “Omics” data, allows protein biologists to study biological and biomedical problems at a new level, and opens new opportunities for both basic research and translational medicine. In this thesis, we look at three closely related topics in computational biology studies of proteins that are deeply influenced by this new technology: Protein evolution, GWAS studies, and protein structure modeling.

### 1.1: Human single nucleotide variants and evolutionary biology

Since the invention of protein sequencing methods in the late 1940s (Edman 1949), multi-species sequence analyses have been applied to study many questions in evolutionary biology. These methods have led to many significant insights and results, for example the construction of a tree of life (Sugden, Jasny et al. 2003), the discovery of lateral gene transfer (Koonin, Makarova et al. 2001), and identification of possible mechanisms for the generation of new proteins (Hughes 1994; Pegueroles, Laurie et al. 2013). On the other hand, population genetics has used theoretical approaches (Crow and Kimura 1970; Ewens 1990) and computer simulations (Schneider, Roessli et al. 2000; Anderson, Ramakrishnan et al. 2005) extensively to

explain various observations, such as the increase in gene number (Lynch and Conery 2003) and expansion in the size and number of intragenic spacers (Lynch and Conery 2003), as well as dramatic proliferation of mobile genetic elements in multicellular eukaryotic organisms (Lynch and Conery 2003; Lynch 2007), based on the genetic drift and selection models (Hartl and Clark 1997).

There is a long history of using experimental populations to demonstrate simple population genetics rules. For example, Buri (Buri 1956) tested the process of allele fixation and elimination under genetic drift through the construction and maintenance of experimental populations of *Drosophila* for 19 generations. That work found the fixation and elimination of eye color mutations, which are considered neutral, followed the predictions of the genetic drift theory. With the development of sequencing technology, it became possible to study population genetics using high quality sequence data on a large scale. For example, a long term *E.coli* evolution experiment (Blount, Borland et al. 2008), started in 1988, has illuminated the genetic basis and evolutionary process underlying the Cit<sup>+</sup> trait (ability to grow on citrate under oxygen-rich conditions) in unprecedented detail (Blount, Barrick et al. 2012). Such direct observation of the adaptation process through sequencing of multiple generations of *E.coli* populations provides new insight into how biological functions emerge.

The recent breakthrough in next generation sequencing technology (Bentley, Balasubramanian et al. 2008; Eid, Fehr et al. 2009; Rothberg, Hinz et al. 2011) makes

large scale genotyping of human genomes possible. The 1000 genomes project (Durbin, Altshuler et al. 2010) provides us with more than 1000 complete genomes from 14 populations of different ethnic backgrounds. This resource captures up to 98% of accessible Human single nucleotide variations (SNVs) at a frequency of 1% or higher and so allows us to study a series of population genomics problems related to protein sequence evolution.

In the first project of this thesis, we address several questions about protein sequence evolution using the SNV data from the 1000 genome project: (1) How does the fitness impact of non-synonymous SNVs affect their population frequency distribution? (2) What is the relationship between the distribution pattern of non-synonymous, synonymous, and intron SNVs for the same protein? (3) How do the factors that are most strongly coupled with protein evolutionary rate such as mRNA expression level (Drummond and Wilke 2008), influence the distribution of human variants?

Our study shows that population allele frequency of a SNV correlates negatively with its impact, as expected from the theory of purifying selection (Hughes, Packer et al. 2003; Bustamante, Fledel-Alon et al. 2005). Synonymous SNV density and non-synonymous SNV density correlate with each other, and they both correlate strongly with intron SNV density. This is compatible with the correlation between  $dN$  (protein non-synonymous evolutionary rate) and  $dS$  (protein synonymous evolutionary rate) discovered in both bacteria and mammals (Li, Wu et al. 1985; Sharp and Li 1987) (Wolfe and Sharp 1993). mRNA expression level is negatively correlated with non-

synonymous SNV density but not synonymous SNV density, compatible with the well-known correlation between mRNA expression level and protein evolutionary rate. In summary, our studies demonstrate that high throughput sequencing data can supplement multi-species protein sequence comparison studies to help us understand the protein sequence evolution from a different perspective.

### 1.2: GWAS and drug targets

In addition to evolutionary biology, another field that benefits greatly from the new sequencing technology is that of GWA studies. GWAS (Genome Wide Association Study) is a method to examine many genetic variants and to determine whether the presence or absence of any variants is associated with a trait, usually a complex disease. The studies usually compare genetic markers or DNA sequences collected from two groups: people with the trait (case) and people without the trait (control) and detect those genetic markers, usually SNPs (Single Nucleotide Polymorphisms), that have significantly different frequency in the case and control groups. There have been thousands of GWA studies for different complex diseases in different populations. The GWAS catalog (<http://www.genome.gov/gwastudies/>) is a central depository for results of GWAS studies.

In a typical GWA study, DNA samples extracted from a few hundreds to thousands of case and control people are genotyped using DNA chip technology, which is a fast and affordable method for simultaneously genotyping 500,000 to one million Single

Nucleotide Polymorphism (SNPs) for a DNA sample (DiStefano and Taverna 2011). The SNPs on a chip (tag SNPs) are a representative subset that are correlated with other SNPs through Linkage Disequilibrium (Jorde 1995) (LD), thus almost every common SNP has a closely correlated tag SNP (Ohashi and Tokunaga 2001). Detection of SNP-phenotype associations requires a large sample size. Meta-studies, the practice of using data from multiple GWA Studies of the same disease together, can increase the effective population size, thus increasing the statistical power to detect weak associations (Zeggini and Ioannidis 2009; Thompson, Attia et al. 2011). Meta-studies have been used for a number of complex diseases and have identified new loci not found by the regular GWAS method (Franke, McGovern et al. 2010; Stahl, Raychaudhuri et al. 2010).

The second project – GWAS identified genes and Drug Targets, is motivated by the puzzling observation that GWA studies do not identify the most important genes for blood pressure regulation. There is a long history of identification of genes affecting blood pressure using non-genomic methods, and 30 genes discovered in this way have provided successful targets for treating hypertension (Johnson, Newton-Cheh et al. 2011). A study has found that among 14 GWAS blood pressure associated genes, only two are in a previous list of 160 traditional candidates (Ehret 2010) and none are established drug targets. This led us to examine whether this phenomenon is common for other diseases. We compared a set of 1621 reported mechanism genes in the GWAS catalog ([www.genome.gov/gwastudies](http://www.genome.gov/gwastudies) (Hindorff, Sethupathy et al. 2009), January 2012) with a corresponding set of 856 known drug target genes (obtained

from Drugbank (Knox, Law et al. 2011), January 2012) for the same diseases. If drug targets were found by GWAS, there should be a large overlap of these two sets. In fact, only 20 of the 856 drug target genes are discovered in GWAS studies of the same disease. We showed that a possible explanation for this is that the drug targets are more tightly coupled with the disease phenotype, thus have fewer variants than are detectable by GWAS. We then look at the relationship between GWAS reported genes and corresponding drug targets in a protein interaction network (Wu, Feng et al. 2010). The network is a protein functional interaction network generated by extending curated biological pathways with non-curated sources of information, including protein-protein interactions, gene co-expression, protein domain interaction, Gene Ontology (GO) annotations and text-mined protein interactions, and covers about 50% of human genes. We found GWAS genes and drug targets for a specific disease are closely related to each other in the network. We exploit that finding to develop a protein network based method for predicting potential drug targets from GWAS identified genes, and as a result were able to propose several possibilities for drug repurposing.

### 1.3: Protein structure modeling

To fully take advantage of the newly generated human SNV data and GWAS results, it's important to measure the biological impact of SNVs. For non-synonymous SNVs, the best way to look at impact is to put them in the context of protein three dimensional structures. At present, the application of structure based methods is

limited by the low experimental coverage of human proteins: for example, currently, only around 8% of the sites of common missense single nucleotide variants lie within an experimentally determined structure of the corresponding protein. In order to analyze a larger proportion of non-synonymous SNVs, it is necessary to extend the experimental coverage using models of structure.

Protein structure modeling is the practice of predicting protein three-dimensional structure based on protein sequence. Generally speaking, there are two kinds of structure modeling methods: *de novo* methods (Lee, Liwo et al. 1999; Rohl and Baker 2002; Chikenji, Fujitsuka et al. 2003; Bradley, Misura et al. 2005; Floudas, Fung et al. 2006; Hung, Ngan et al. 2007) and homology modeling (Sali and Blundell 1994; Cardozo, Totrov et al. 1995; Bower, Cohen et al. 1997; Bordoli, Kiefer et al. 2008; Moult 2008). *De novo* methods predict the protein structure from physics and chemistry principles without any template. On the other hand homology modeling predicts the structure of a target protein based on the known structure of a closely related template protein. The quality of homology modeling depends on the sequence identity between the template and the target. Models of different qualities have different applications (Moult 2008). High quality models are useful in fields that require high accuracy such as drug design, while models of low quality are often useful in gene function annotation, for example. For the purpose of predicting the impact of human missense single-nucleotide variants (SNVs), intermediate quality models are required, and it's usually possible to make a reasonable prediction using models built on sequence identity of above 40%. We found the main source of errors

in SNV impact prediction using lower sequence identity models is mis-alignment of the target and template sequences, rather than long term structural divergence. In the third project, we therefore designed a method to filter out models built with bad alignments and use only those with reliable ones. We developed a dynamic programming based algorithm to evaluate the quality of pairwise protein sequence alignments and demonstrated that we can model more non-synonymous variants and analyze their impact with reasonable accuracy.

#### 1.4: Overview

The dissertation is organized as follows. Chapter 2 first discusses the relationship between evolutionary biology and population genomics, then goes through a series of topics on the distribution pattern of human variants, discussing their implications in protein evolution. Chapter 3 discusses the finding that GWAS usually does not identify genes most tightly coupled to the disease mechanism, and proposes an explanation for it. Then we describe a protein network based method to overcome this issue. Chapter 4 introduces a new algorithm to evaluate the accuracy of pairwise protein sequence alignment, and then demonstrates its application in human missense SNV impact modeling. Chapter 5 summarized the conclusion of the three projects and discussed future perspectives in these areas.

## Chapter 2: Distribution of human single nucleotide variants and implications for protein sequence evolution

### 2.1: Introduction

Protein sequence evolution is a central topic in biology. Knowledge of this process can help us understand the process of speciation (Webster, Payne et al. 2003), facilitate the identification of functionally important sites in proteins (Reetz, Wang et al. 2006), find protein interaction partners (Ramani and Marcotte 2003), predict protein structure (Hopf, Colwell et al. 2012), and predict the impact of non-synonymous variants in human disease (Ng and Henikoff 2003; Yue, Melamud et al. 2006). Inter-species sequence analyses have revealed many features of protein sequence evolution. For example, positive correlations between dN, non-synonymous mutation rate, and dS, synonymous mutation rate, have been discovered in both mammals (Wolfe and Sharp 1993) and bacteria (Sharp and Li 1987). Combined analysis of inter-species protein sequence alignment and large-scale biological experimental data have identified several factors that significantly influence the rate of protein sequence evolution, such as mRNA expression level (Drummond and Wilke 2008) and the number of protein-protein interactions (Pal, Papp et al. 2006). With the development of next generation sequencing technology (Bentley, Balasubramanian et al. 2008; Eid, Fehr et al. 2009; Rothberg, Hinz et al. 2011), we are accumulating a large amount of human individual sequence data. These data

allow us to revisit these observations from the perspective of intra-species sequence analysis.

There is a long history of population genetics studies of the evolution process through intra-species sequence comparison. In particular, a number of statistical measures have been developed for assessing deviations from neutrality. For example, Tajima's  $D$  is a measure based on the difference between the observed number of pairwise segregating sites and the number of pairwise segregating sites expected from the total polymorphic sites for a group of sequences based on a neutral model (Tajima 1989). Other widely used statistics include Fu's  $F$  (Fu 1995), Fay Wu's  $H$  (Fay and Wu 2000), and Zeng's  $E$  (Zeng, Fu et al. 2006) are all based on the difference between some observed quantity and the expectation for that quantity from the assumption of neutrality. These measures have been used in studies for model organisms (Bachtrog 2004), agriculture plants (Wright, Bi et al. 2005), pathogens (Shriner, Shankarappa et al. 2004) and human populations. Studies range from small scale (a couple of genes) (Polley and Conway 2001) to a moderately large scale (hundreds of genes) (Stephens, Schneider et al. 2001; Akey, Eberle et al. 2004), with the general goal of finding genes or more broadly speaking regions of DNA that deviate from neutrality. The results have provided extensive information on regions undergoing both negative and positive selection, as well some artificial selection (Wright, Bi et al. 2005) for some plants.

A more relevant population genetics measure for studies of protein sequence

evolution is the McDonald Kreitman test (McDonald and Kreitman 1991). This test is designed to detect positive selection based on the contrast between the inter-species divergence pattern and the intra-species polymorphic pattern, comparing the ratio between the number of polymorphic and fixed sites for synonymous mutations with that for non-synonymous mutations. The test has been applied to many model organisms (Nachman, Boyer et al. 1994; Smith and Eyre-Walker 2002) and has provided information on the extent of adaptive evolution between species, showing this to vary widely over mammals, *Drosophila*, plants and bacteria (Eyre-Walker 2006).

A suitable measure for specifically studying variation within a population and its relationship to other factors is single nucleotide variant (SNV) density. For example, early studies (Zhao, Fu et al. 2003) calculated the SNP (Single Nucleotide Polymorphism) density for different categories of human SNPs, (Chen and Rajewsky 2006) used SNP density to measure selection on miRNA binding sites. (Gerstein, Kundaje et al. 2012) used non-synonymous SNV density together with dN/dS values to measure the selection pressure for different genes in the analysis of a human regulatory network constructed from ENCODE project data (Ecker, Bickmore et al. 2012). At the population level, non-synonymous SNV density,  $N_d$ , reflects the tolerance of a gene to non-synonymous mutations, and so can be considered a population level measure of dN. Similarly the synonymous SNV density,  $S_d$ , can be considered a population level measure of dS. There is an important difference between these two sets of measures (Kryazhimskiy and Plotkin 2008): dN and dS

measure the rate of substitutions along divergent lineages—i.e., the rate at which mutations arise and subsequently fix; whereas the two densities focus on the behavior of segregating mutations in the process of elimination and fixation. Nevertheless, they are all directly related to protein sequence evolution and give us measures of the variability of protein sequence at two different levels.

One of the disadvantages of this type of analysis is that the density treats every variant equally, regardless of population frequency. The population frequency of a variant in part reflects the selection pressure acting on it, so that common variants are more likely to be neutral than rare ones (Hughes, Packer et al. 2003; Bustamante, Fledel-Alon et al. 2005). The other disadvantage is that the approach treats every variant independently, thus neglecting the fine structure of linkage between variants. Previous studies have shown that there is significant hitchhiking of variants in coding and regulatory sites in human (Cai, Macpherson et al. 2009), so the density is to some extent distorted by such effects. Both factors may reduce but not invalidate the signals caused by the functional factors that protein evolutionary biologists are interested in. The advantage of the density measure is its relevance to the multi-species protein sequence analysis measures such as  $dN$  and  $dS$ . That relationship allows us to translate certain protein sequence evolution problems previously studied at the multi-species level into the language of population genomics, and thus study these problems from that perspective.

The 1000 genomes project (Durbin, Altshuler et al. 2010), an international effort to

characterize the geographic and functional spectrum of human genetic variation, has provided us with more than 1000 complete genomes from 14 populations of different ethnic backgrounds. It allows us to study a lot of human population genomics problems which are impossible before. In this project, we first look at the relationship between the impact of non-synonymous single nucleotide variants (SNVs) and population allele frequencies, and show a significant negative correlation between these quantities, as is expected from the theory of purifying selection (Hughes, Packer et al. 2003; Bustamante, Fledel-Alon et al. 2005). We then revisited the problem of correlation between dN and dS using these SNV densities and found significant correlation between non-synonymous variant density and synonymous variant density, consistent with the previously observed correlation between dN and dS (Sharp and Li 1987; Wolfe and Sharp 1993). The correlation suggests non-synonymous and synonymous sites may be under the influence of some common factors. An alternative explanation for the correlation between dN and dS is that it reflects the fluctuation of local mutation prevalence (Ohta and Ina 1995), so that dN and dS are both high in high mutation prevalence regions of the genome and both low in low mutation prevalence regions. Studies have found that the variability of mutation prevalence across the genome can be ascribed to unexpected factors such as distance to nearby insertions and deletions (Tian, Wang et al. 2008). Later studies of the same subject argued that the increased mutation prevalence near indels is caused by nearby repeated sequences, which promotes an increased probability of replication fork arrest, causing the persistent recruitment of error-prone DNA polymerases (McDonald, Wang et al. 2011). It's still under debate whether the variation of

mutation prevalence across mammalian genomes is adaptive or not (Hodgkinson and Eyre-Walker 2011). In *C.elegans*, it has been found that the rate of spontaneous single-base mutation is uniform on a chromosome scale (Denver, Dolan et al. 2009) and the observed prevalence of variations is dominated by background selection: There is a reduction in neutral variation due to linkage between neutral variants and deleterious mutations undergoing elimination from the population (Hudson and Kaplan 1995), a process highly dependent on the local recombination rate (Rockman, Skrovanek et al. 2010) . In this study, we measure the local mutation prevalence through the variant density of introns.

Introns comprise 37% of the human genome (Fedorova and Fedorov 2005). SNVs in introns do not change coding sequence or codon usage, and are thus unlikely to have any impact on protein structure and function, or on translation efficiency. Introns may be involved in regulation of gene expression (Ratajewski, de Boussac et al. 2012) and mutations in introns may cause splicing errors (Kulseth, Berge et al. 2010). The functional roles of introns and their importance to the fitness of individuals are under debate (Parenteau, Durand et al. 2008; Melamud and Moulton 2009). Whatever functional roles may eventually become clear, it is already evident that most of the bases in these regions are not involved in function, and so intron variation provides a useful tool for measuring the local mutation prevalence of DNA. For example, (Metz, Robles-Sikisaka et al. 1998) used the substitution rate in introns as a reference to detect positive selection in abalone sperm fertilization genes. (Parsch, Novozhilov et al. 2010) used the intron mutation rate as a reference to detect selection in *Drosophila*

genes. We calculated the SNV density for introns using the 1000 genome data and found significant correlation between this quantity with both the density of non-synonymous SNVs and the density of synonymous SNVs, consistent with variation in mutation prevalence across the genome driving the correlation of non-synonymous and synonymous SNV densities. This observation suggests the correlation between non-synonymous SNV density and synonymous SNV density could be the byproduct of their common correlation with the SNV density of introns, which is a measure of local mutation prevalence. We used partial correlation analysis to look at the correlation between synonymous SNV density and non-synonymous SNV density with the effect of intron SNV density removed. Partial correlation (Johnson and Wichern 2002; Baba, Shibata et al. 2004) is a statistical procedure to remove the effect of a third random variable when calculating the correlation between two random variables. The analysis showed that the correlation between density of synonymous SNV and density of non-synonymous SNV can't be fully explained by their common correlation with intron SNV density.

Studies have shown that an unexpected factor, mRNA expression level, is the most significant determinant of evolutionary clock rate. significant negative correlations between protein sequence divergence rate and mRNA expression level have been found in bacteria (Rocha and Danchin 2004), yeast (Drummond, Raval et al. 2006), worm (Krylov, Wolf et al. 2003), plants (Wright, Yau et al. 2004), fruit flies (Lemos, Bettencourt et al. 2005) and humans (Subramanian and Kumar 2004). Drummond and Wilke proposed a model to explain this striking correlation (Drummond and Wilke

2008) in terms of selection against misfolded proteins: Mutations in proteins lead to protein misfolding, leaving exposed hydrophobic residues. These residues will bind to hydrophobic sites on other proteins, leading to protein aggregation, which has serious consequences for the fitness of the cell (Bucciantini, Giannoni et al. 2002). Drummond and Wilke argued that this effect increases with expression level leading to the observed negative correlation between mRNA expression level and protein sequence evolutionary rate. These analyses are all based on inter-species protein sequence comparisons to determine rates. Here we look at the problem from population genomics perspective by examining the relationship between non-synonymous SNV density and mRNA expression level. We observed that non-synonymous SNV density negatively correlates with mRNA expression level. We did not observe significant correlation between synonymous SNV density and mRNA expression level, inconsistent with the inter-species results. We also looked at the correlation between non-synonymous variant density and the number of protein-protein interactions. A previous study showed that the number of protein-protein interactions is also a determinant of protein evolutionary rate, though not as strongly as mRNA expression level (Pal, Papp et al. 2006). Some studies suggest that only the most prolific interactors tend to evolve slowly (Jordan, Wolf et al. 2003), but other studies suggests that the weaker overall correlation is the result of incomplete interaction data, and once a complete interactome is included, the correlation is significant (Fraser, Wall et al. 2003) . At the population level, we found significant correlation between non-synonymous SNV density and the number of protein-protein interactions. Using partial correlation analysis (Johnson and Wichern 2002), we found

the influence of number of protein-protein interactions on the non-synonymous SNV density is independent of the influence of mRNA expression level.

Finally, we look at genes that deviate from these general trends and examined the functions of these “outlier” genes, discovering interesting patterns for them.

## 2.2: Results

### **Deleterious variants occur at lower frequencies than neutral variants.**

Deleterious SNVs are under purifying selection, so it's expected that their frequencies would be lower than those of neutral SNVs. We use the SNPs3D profile method (Yue, Melamud et al. 2006) to calculate the expected impact on *in vivo* protein function of the non-synonymous SNVs in the 1000 genomes data. The SNPs3D profile method uses a support vector machine (SVM) which returns a score related to the probability that a SNV is deleterious. The more negative the score, the more probable that SNV is deleterious. We found the fraction of predicted deleterious SNVs decreases with increasing allele frequency. This is consistent with the population genetics model: Neutral variants are more likely to reach high frequency and even fixation in the population while deleterious variants are more likely to be eliminated, though their chance of fixation is not zero (Hartl and Clark 1997). Similar results are also found in (Marth, Yu et al. 2011).

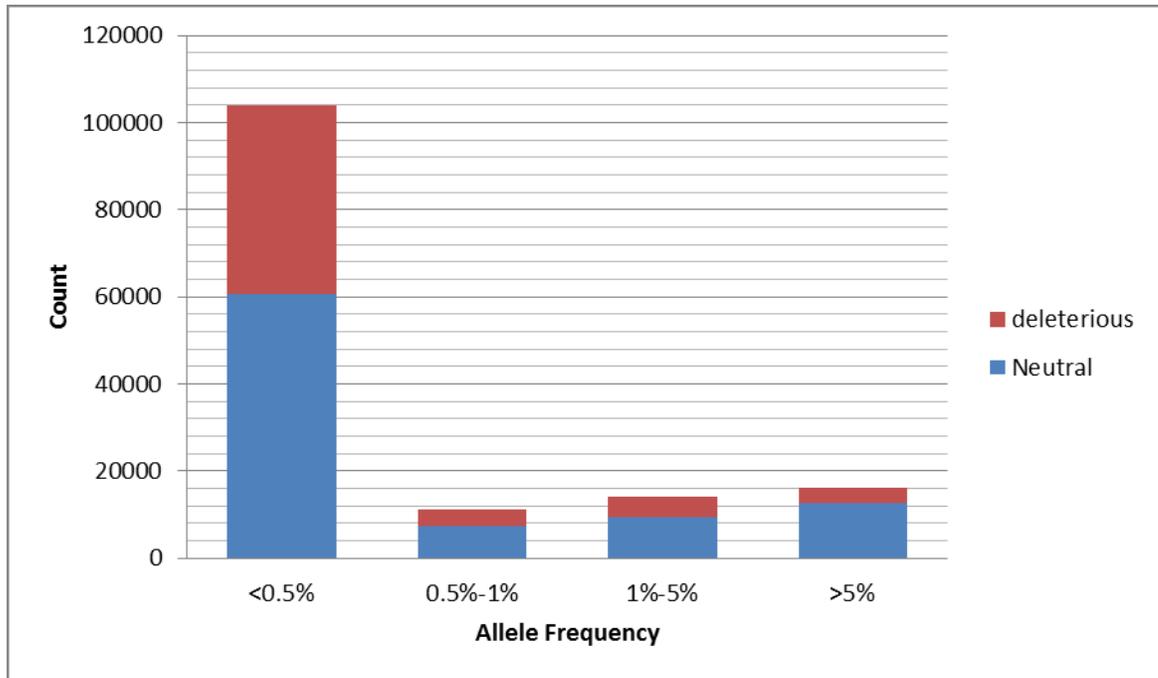


Figure 2.1 Number of predicted deleterious and neutral single ns-SNVs in the 1000 genomes data as a function of allele frequency. Most alleles occur at a frequency of less than 0.5%. The fraction of SNVs that are predicted deleterious is much higher (approaching 50%) in the low frequency SNVs. Above 5% frequency the fraction predicted deleterious is 22.5%.

### **Correlation between non-synonymous and synonymous variant density**

We observed a substantial positive correlation between synonymous and non-synonymous SNV densities (Figure 2.2A, 2.2B), consistent with the many

observations of a correlation between dN and dS. As noted earlier, there are two suggested explanations for these correlations – selection for codon use and variation in local mutation prevalence across the genome. Rates of sequence change in introns can in principle provide a useful means of distinguishing between these possibilities. If the cause is variation in local mutation rate there will be a correlation between the intron rate of change and coding related rates. But due to varying intron structures between species, the calculation of substitution rate in these regions is not possible except for closely related species such as Human and chimpanzee (Gazave, Marqués-Bonet et al. 2007). An advantage of population data is that we can calculate SNV density,  $I_d$ , in introns, providing a good measure of local mutation prevalence. We find a correlation between intron SNV density and both non-synonymous SNV density (Correlation: 0.37,  $p < 2.2e-16$ ) and synonymous SNV density (Correlation: 0.39,  $p < 2.2e-16$ ) (Figure 2.2C, 2.2D). Thus there are correlations between all pairs of densities:  $N_d$  with  $S_d$ ,  $N_d$  with  $I_d$ , and  $S_d$  with  $I_d$ . To test whether the correlation between synonymous SNV density and non-synonymous SNV density is the consequence of regional mutation prevalence variation, we calculated partial correlation (Baba, Shibata et al. 2004) between non-synonymous SNV density and synonymous SNV density conditional on the intron SNV density. We found that the partial correlation is still very significant (Correlation: 0.18,  $P < 2.2e-16$ ) after removing the influence of intron variant density, suggesting that regional mutation prevalence fluctuation can only explain part of the correlation between synonymous and non-synonymous SNV density. So the correlation between  $N_d$  and  $S_d$  is more likely to be the combined effect of shared selection pressure between non-

synonymous and synonymous sites and local mutation prevalence variation.

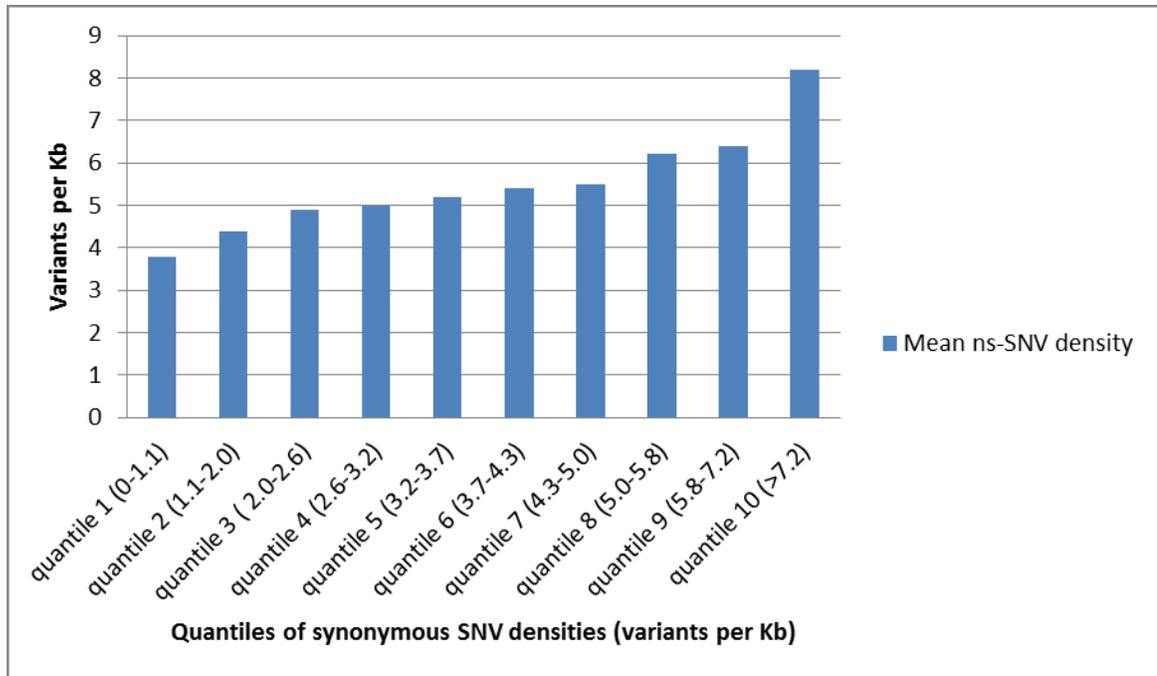


Figure 2.2A Mean ns-SNV densities for different quantiles of synonymous SNV density. The correlation coefficient between ns-SNV density and synonymous SNV density is 0.30 ( $P < 2.2e-16$ ).

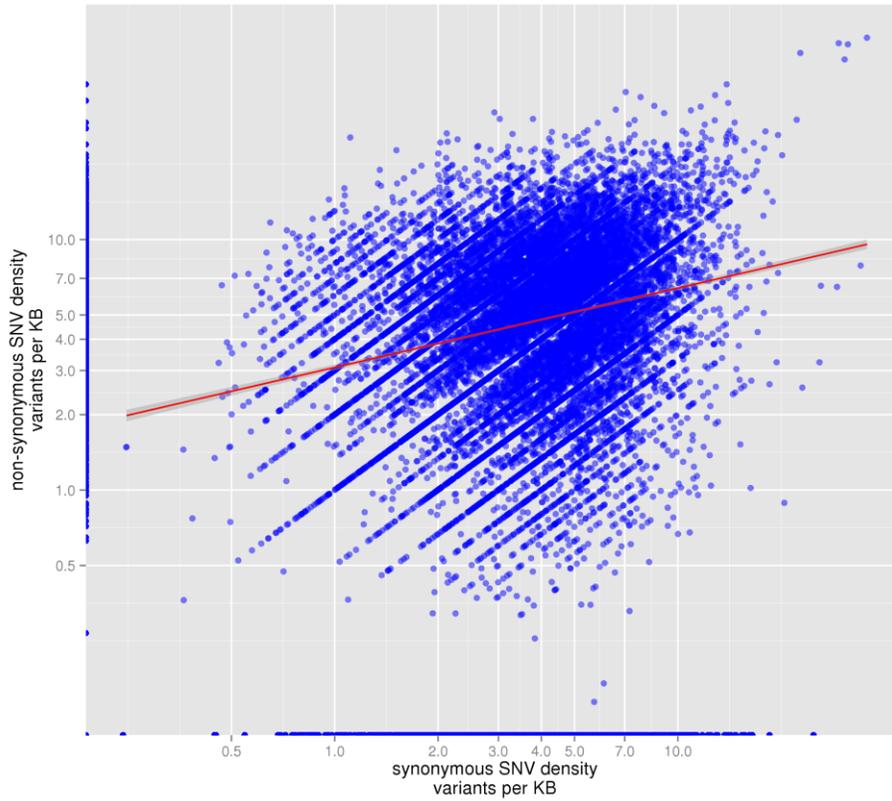


Figure 2.2B Dot plot of ns-SNV density versus synonymous SNV density in log scale for all genes. The red line is smooth line generated using a linear model. The correlation coefficient between ns-SNV density and synonymous SNV density is 0.30 ( $P < 2.2e-16$ ).

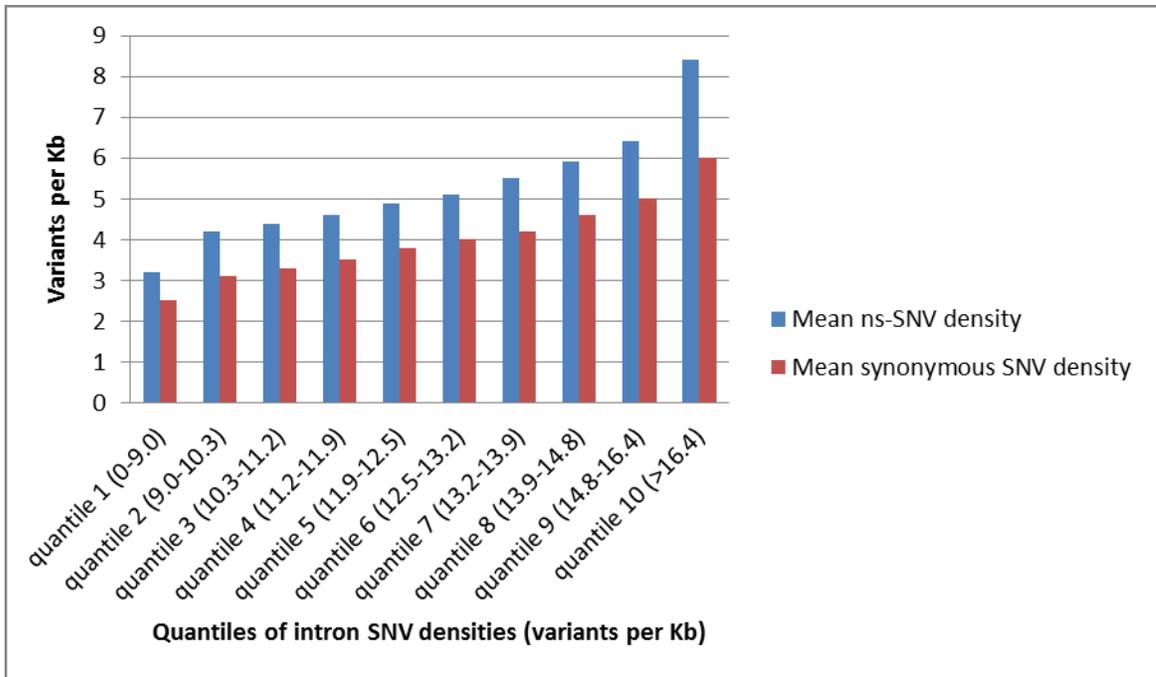


Figure 2.2C Distribution of the mean ns-SNV density and synonymous SNV density for different quantiles of intron SNV density. From this distribution we can observe correlation between both ns-SNV density and synonymous SNV density with intron SNV density. The correlation coefficient between ns-SNV and intron SNV densities is 0.37 ( $P < 2.2e-16$ ) and the correlation coefficient between synonymous-SNV and intron SNV densities is 0.39 ( $P < 2.2e-16$ )

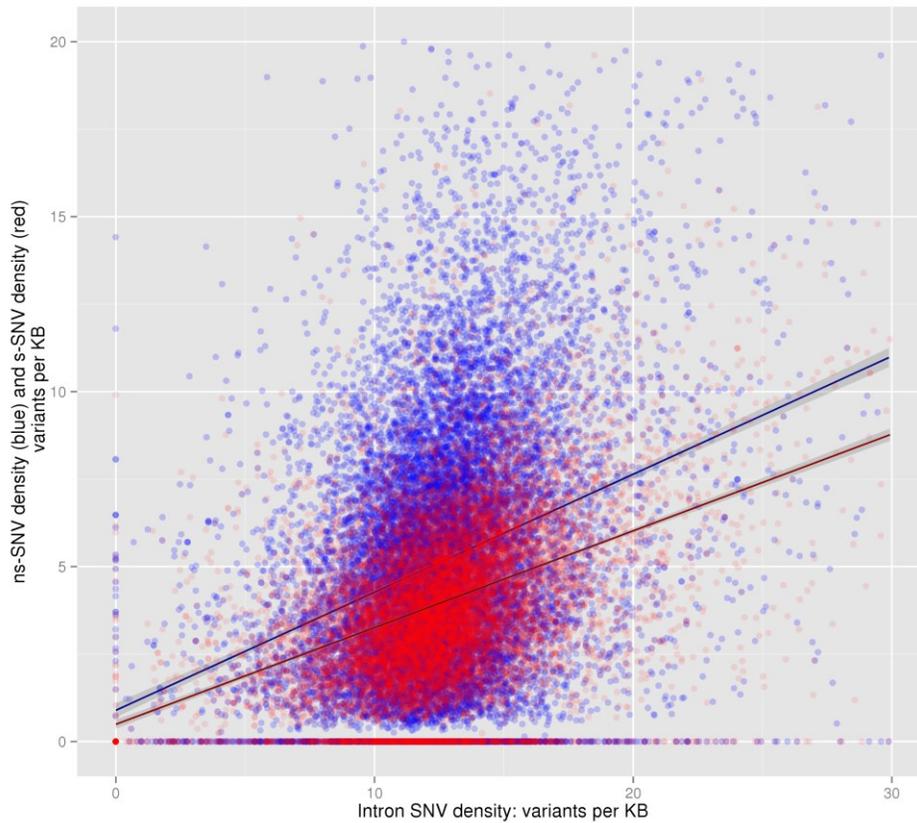


Figure 2.2D Dot plot of ns-SNV density (blue) and synonymous SNV density (red) versus intron SNV density for all genes. The dark blue line is the trend line for the relationship between ns-SNV density and intron SNV density generated by a linear model, the dark red line is the trend line for the relationship between synonymous SNV density and intron SNV density generated by a linear model. The x-axis is truncated at 30 per KB and the y axis is truncated at 20 per KB. The correlation coefficient between ns-SNV and intron SNV densities is 0.37 ( $P < 2.2e-16$ ) and the correlation coefficient between synonymous-SNV and intron SNV densities is 0.39 ( $P < 2.2e-16$ )

### **ns-SNV densities are restrained by the mRNA expression level**

We examined the most significant determinant of protein evolutionary rate—mRNA expression level, to see how it influences the process of non-synonymous variant elimination and fixation.

Figure 2.3 shows the relationship between expression level and the density of non-synonymous and synonymous SNPs. The observed correlation of non-synonymous SNV density with expression level is compatible with what various studies have found from protein sequence analysis (Krylov, Wolf et al. 2003; Rocha and Danchin 2004; Subramanian and Kumar 2004; Wright, Yau et al. 2004; Lemos, Bettencourt et al. 2005; Drummond, Raval et al. 2006). The population genomics data not only provides new supporting evidence that mRNA expression level is a strong determinant of selection pressure for protein evolution, but also gives us a way to estimate the exact level of constraint caused by expression level. Using a simple linear model, we estimate that non-synonymous SNV density will decrease by 0.18/Kb (see Methods) for each doubling of the expression level, which is roughly 1/25 of the average ns-SNV density for all genes. We found there is no significant correlation between synonymous SNV density and mRNA expression level which is at odds with result of a previous study (Drummond and Wilke 2008).

As noted earlier, according to the model of “mistranslation induced mis-folding” (Drummond and Wilke 2008), mutations cause protein misfolding, and hydrophobic residues exposed as a consequence seek hydrophobic partners in other misfolded

proteins, resulting in aggregation that disrupts cell function. If this mechanism is the main selection pressure against the accumulation of non-synonymous changes, we expect that the trend of negative correlation between the SNV density of high impact variants and mRNA expression will be larger than that between the SNV density of low impact variants and mRNA expression. We tested this by dividing non-synonymous SNVs into two categories: those with an expected high impact on protein function and those expected to be neutral by our SNPs3D profile method. We found that the correlation to mRNA expression level is slightly stronger for predicted high impact SNVs than neutral SNVs (Figure 2.4), consistent with the mistranslation induced mis-folding model.

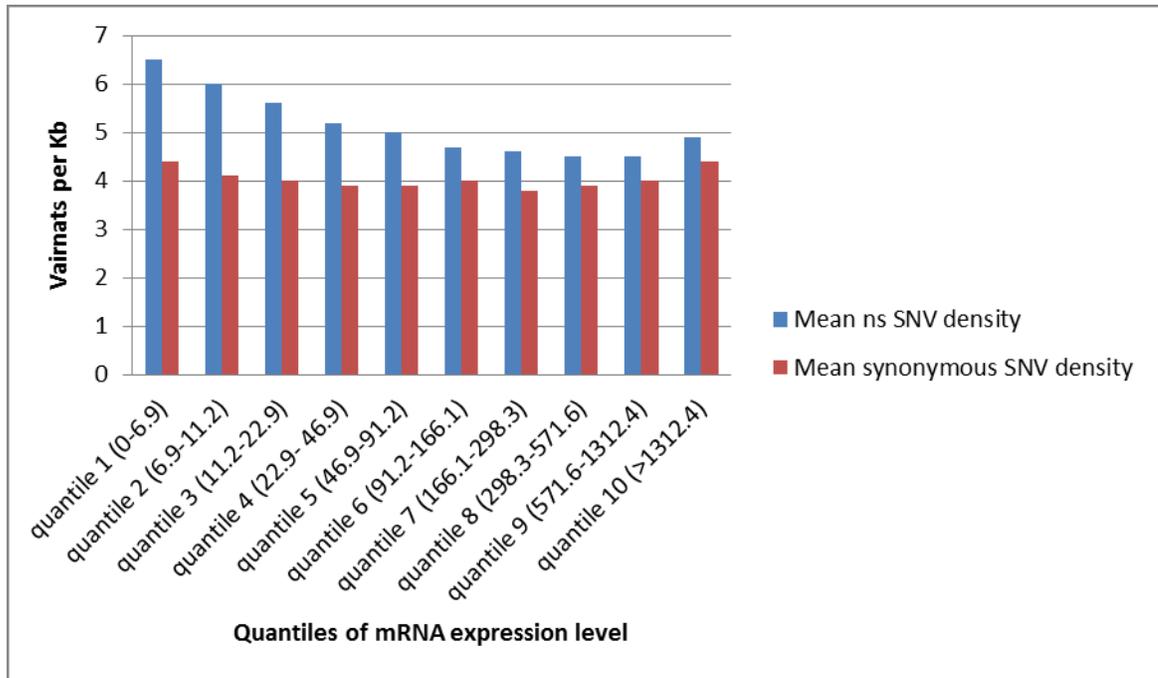


Figure 2.3 Mean ns-SNV and synonymous SNV density for different quantiles of mRNA expression levels. The ns-SNV density negatively correlates with the logarithm of mRNA expression level (Correlation coefficient -0.13,  $P < 2.2e-16$ ). There is no significant correlation between synonymous SNV density and logarithm of mRNA expression level.

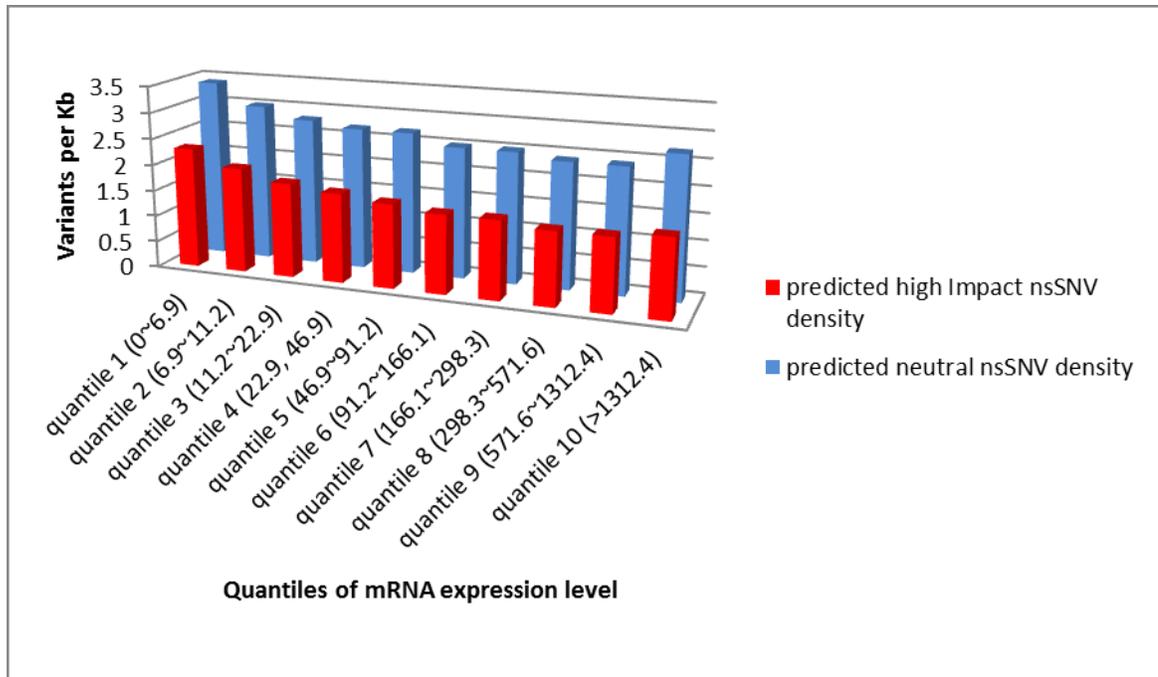


Figure 2.4 Mean ns-SNV density of two different categories of ns-SNV (red: predicted to be deleterious by SNPs3D, blue: predicted to be neutral by SNPs3D) for different quantiles of mRNA expression level. The correlation of log mRNA expression level with predicted high impact nsSNV density (correlation coefficient: -0.12,  $P < 2.2e-16$ ) is higher than the correlation between the predicted neutral nsSNV density (correlation coefficient: -0.08,  $P < 2.2e-16$ )

### ns-SNV densities also correlate with protein network features

Studies have also found a correlation between the number of protein-protein interactions with protein evolutionary rate (Fraser, Hirsh et al. 2002; Fraser, Wall et al. 2003; Pal, Papp et al. 2006). This correlation is believed to be caused by protein functional restraints rather than the biophysical restraint (Drummond and Wilke

2008) that has been proposed as an explanation for mRNA expression level effects. We look at this question at the population level by comparing the SNV densities of genes with different number of interactions in a protein function interaction network (Wu, Feng et al. 2010). We observe a significant negative correlation between non-synonymous SNV density and the logarithm of the degree of proteins in the network (correlation coefficient: -0.12,  $P < 2.2e-16$ ) and no significant correlation between synonymous SNV density and the logarithm of the degree of proteins in the network (correlation coefficient: -0.02,  $P=0.11$ ). Inspection of the mean ns-SNV density at different quantiles of degree (Figure 2.5) shows the first few bars deviating substantially from the correlation trend. This probably reflects the effect of variable levels of missing interactions in the network. There is also a significant correlation between log degree of proteins and the log expression level (correlation coefficient 0.08,  $P=2.6e-13$ ), so it's possible that the correlation between non-synonymous SNV density and the log of degree is a byproduct of the correlation between non-synonymous SNV density and log expression. To resolve this issue, we calculated the partial correlation between log degree and non-synonymous SNV density conditioned on the log of gene expression levels. We found the partial correlation is very significant (correlation coefficient -0.11,  $P = 1.05e-24$ ). We also calculated the partial correlation between non-synonymous SNV density and log expression level conditioned on the log degree and also found a significant correlation (correlation coefficient -0.13,  $P < 1e-30$ ). These results suggest the number of functional interactions and the mRNA expression level are two independent forces in shaping the selection force on non-synonymous mutation sites for genes. We estimate that the

non-synonymous SNV density will drop 0.21 per Kb for each doubling of functional interactions, roughly the same as the drop of ns-SNV density when the expression level is doubled.

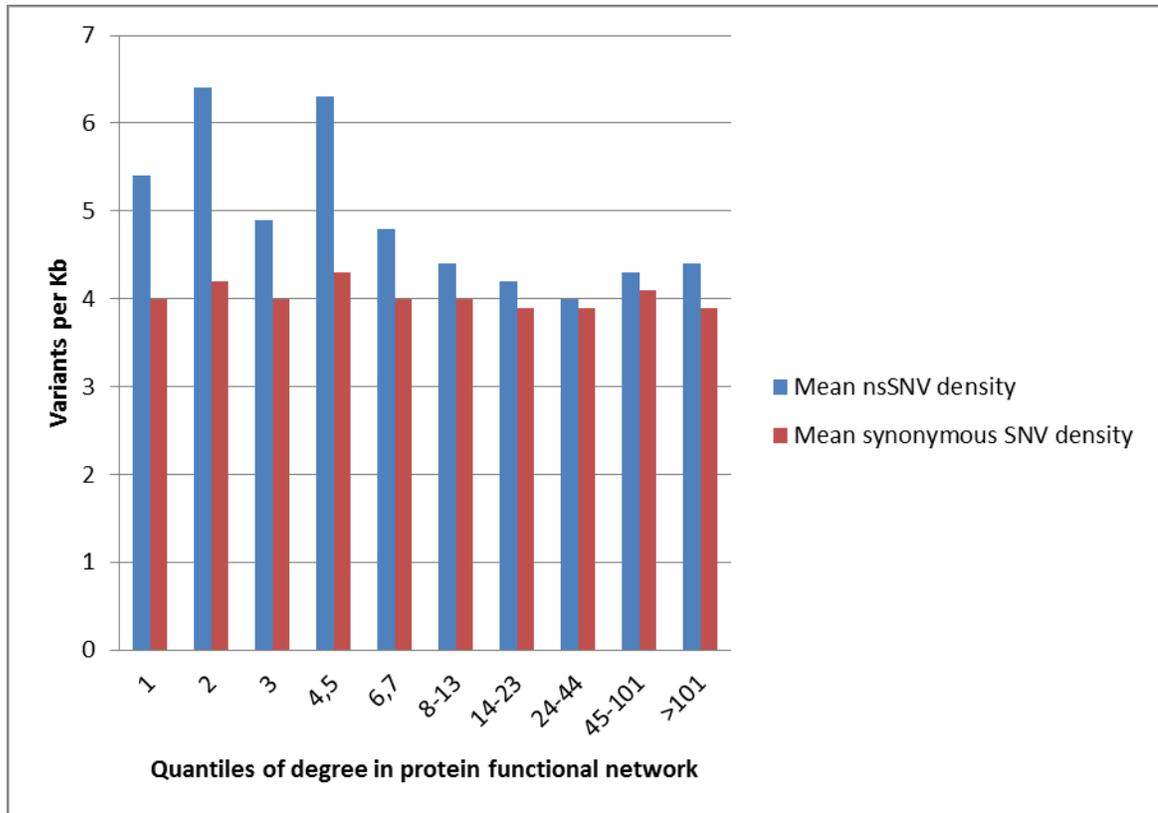


Figure 2.5 Mean ns-SNV/synonymous SNV density as a function of quantiles for the number of protein neighbors in a protein functional network. nsSNV density is negatively correlated with the log degree in the network (correlation coefficient -0.12,  $P < 2.2 \times 10^{-16}$ ), while there is no significant correlation (correlation coefficient -0.02,  $P = 0.11$ ) between synonymous SNV density and the log degree.

## **Outlier genes**

There are some genes that clearly do not follow the general trends described above. It is of interest to examine them. For the correlation between non-synonymous SNV density and mRNA expression level, we pick the 37 genes whose mRNA expression is in the top 10% range and ns-SNV density is at least 3 times the average (Figure 2.6), and perform a GO enrichment test using Gorilla (Eden, Lipson et al. 2007; Eden, Navon et al. 2009). We found these genes are enriched in immune response processes (Table 2.1). Diversity in these genes can increase the fitness of the carrier, offsetting the fitness cost of potential protein aggregation, so it's not surprising that these genes deviate from the general trend. In a similar analysis of non-synonymous SNV density and degree of protein interactions, we pick the 151 genes whose number of interactions is in the top 10% range ns-SNV density is at least 3 times the average. We found a group of genes with degree around 380 having significantly higher non-synonymous variant density than other genes of similar degree (Figure 2.7). Upon inspection of this group, we found they are all zinc finger proteins and that these form an almost complete clique in the functional network. This may be because all these proteins bind DNA so the functional interaction network considers them involving in a super large complex, resulting in an artificial blowup of their degree values. In addition, we found a group of genes with moderately high degree but a very high level of ns-SNV density (Figure 2.7). These are all HLA genes, which have functional interactions with many T-cell receptors. It's not surprising these have high diversity given their role in immune response processes.

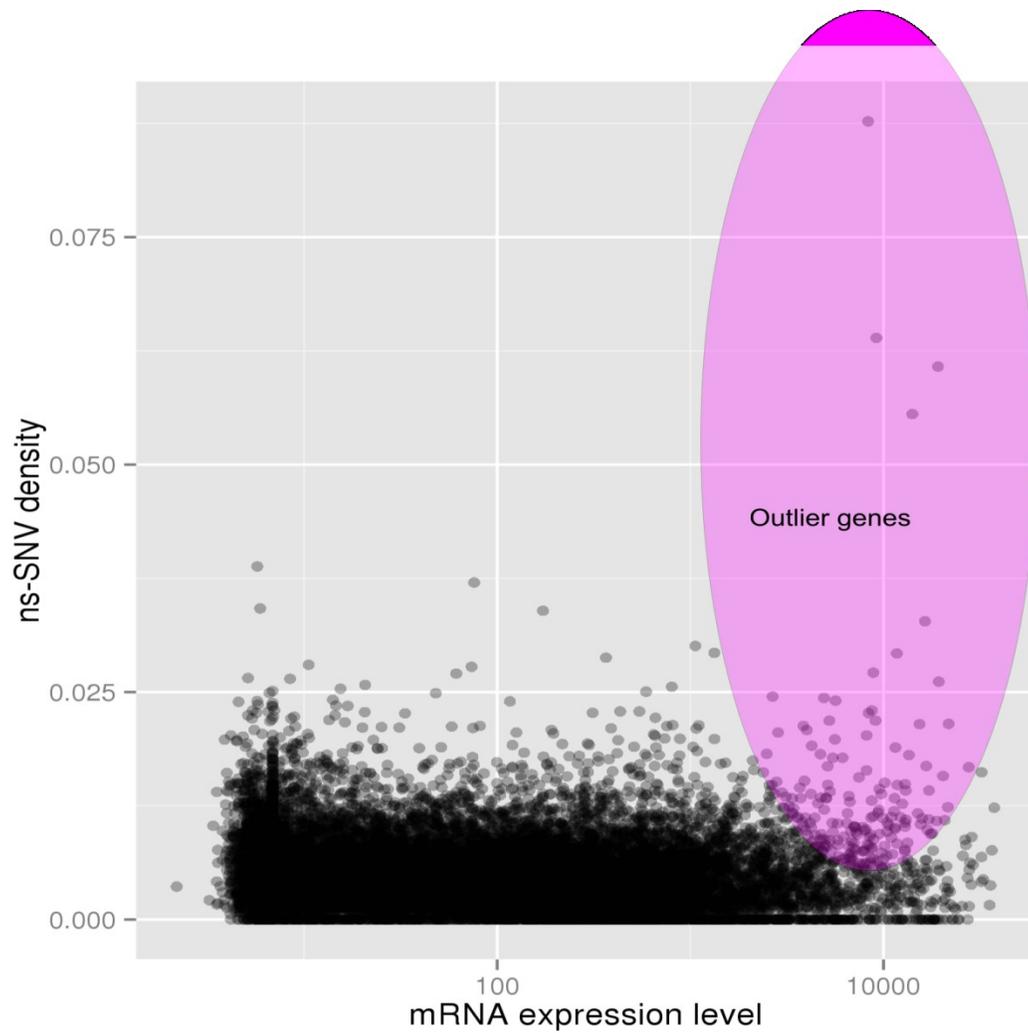


Figure 2.6 Distribution of ns-SNV density versus mRNA expression. Genes in the region highlighted by the oval have very high non-synonymous variant density and high expression values, contrary to the general correlation between these variables. Genes in this region are highly enriched for immune function proteins (see Table 2.1)

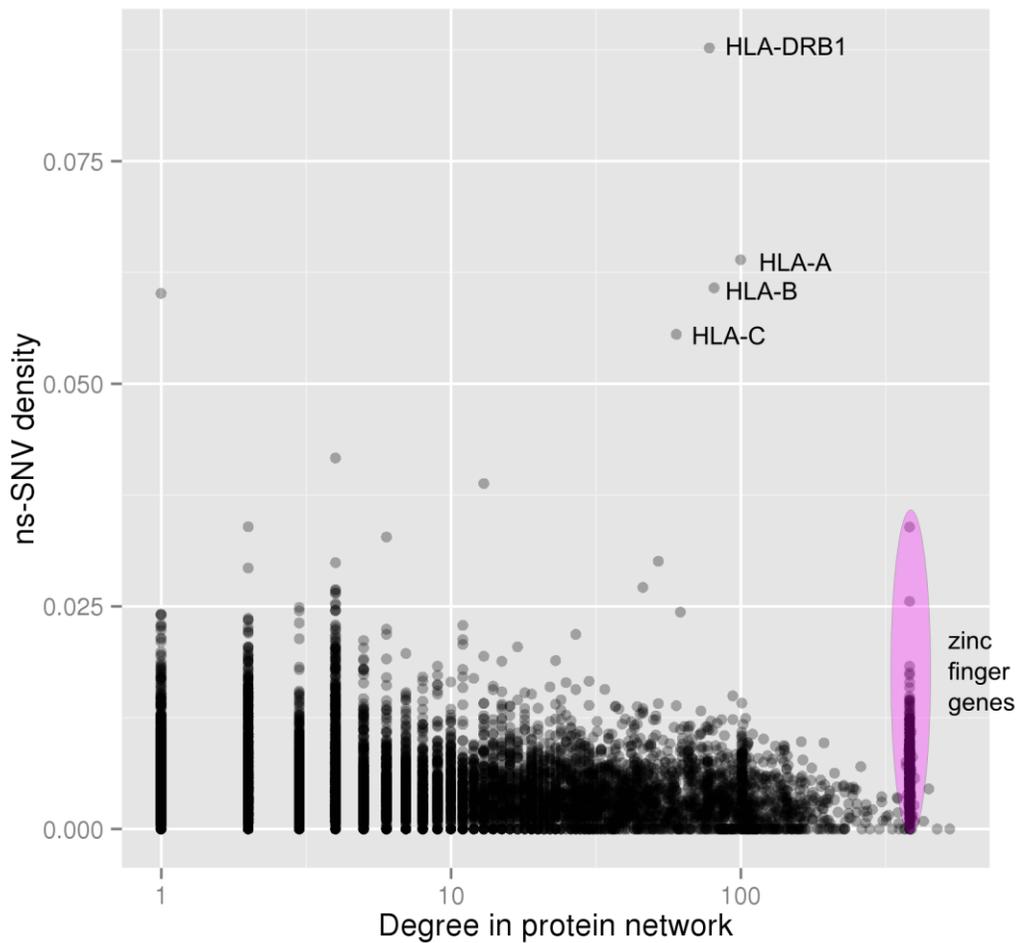


Figure 2.7 Distribution of ns-SNV density versus degree of protein interactions.

Zinc finger genes deviate from the negative correlation between these variables. This is a consequence of the way in which complexes are defined in the interaction network.

Table 2.1 List of GO Biological Process terms enriched in outliers of the correlation between mRNA expression and non-synonymous variant density.

GO Term	Description	P-value	FDR q-value	Enrichment
---------	-------------	---------	-------------	------------

GO:0009617	response to bacterium	2.96E-13	3.34E-09	32.69
GO:0051704	multi-organism process	1.12E-10	6.32E-07	6.45
GO:0051707	response to other organism	3.42E-10	1.29E-06	13.14
GO:0050832	defense response to fungus	6.26E-10	1.77E-06	110.8
GO:0009620	response to fungus	4.73E-09	1.07E-05	76.7
GO:0042742	defense response to bacterium	4.96E-09	9.34E-06	27.64
GO:0060333	interferon-gamma-mediated signaling pathway	8.15E-09	1.32E-05	39.23
GO:0009607	response to biotic stimulus	1.67E-08	2.35E-05	9.07
GO:0044364	disruption of cells of other organism	2.35E-08	2.95E-05	122.73
GO:0031640	killing of cells of other organism	2.35E-08	2.66E-05	122.73
GO:0071346	cellular response to interferon-gamma	3.12E-08	3.21E-05	31.49
GO:0034341	response to interferon-gamma	1.06E-07	9.95E-05	25.73
GO:0001906	cell killing	4.80E-07	4.18E-04	61.36
GO:0002480	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-independent	1.21E-06	9.74E-04	132.95
GO:0006952	defense response	1.65E-06	1.24E-03	5.12
GO:0002478	antigen processing and presentation of exogenous peptide antigen	2.01E-06	1.42E-03	15.64
GO:0019884	antigen processing and presentation of exogenous antigen	2.16E-06	1.44E-03	15.44
GO:0048002	antigen processing and presentation of peptide antigen	3.11E-06	1.95E-03	14.5
GO:0016045	detection of bacterium	4.08E-06	2.43E-03	92.05
GO:0035821	modification of morphology or physiology of other organism	5.06E-06	2.86E-03	34.68
GO:0019882	antigen processing and presentation	6.02E-06	3.24E-03	12.94
GO:0009595	detection of biotic stimulus	2.48E-05	1.27E-02	52.03
GO:0007565	female pregnancy	3.22E-05	1.58E-02	21.86
GO:0019221	cytokine-mediated signaling pathway	6.14E-05	2.89E-02	8.58
GO:0044703	multi-organism reproductive process	7.81E-05	3.53E-02	5.48
GO:0044706	multi-multicellular organism process	1.02E-04	4.44E-02	16.28

### 2.3: Methods

#### **1000 genomes project data**

The 1000 genomes data (Durbin, Altshuler et al. 2010) was downloaded from <http://www.1000genomes.org/data>. The 2010 November data set (Phase I) is used. Only variants that passed all quality filters are accepted for this study. The synonymous, non-synonymous and intron variants are annotated using gene coordinate information from Refseq (Pruitt, Tatusova et al. 2007) genes downloaded from UCSC genome browser (Fujita, Rhead et al. 2011) <http://genome.ucsc.edu/> in Jan 2012. We get 133,103 synonymous, 177,191 non-synonymous and 1,2511,175 intron variants respectively. These numbers are smaller than those in (Abecasis, Auton et al. 2012) because we use Refseq annotation which is more conservative than the gencode annotation (Harrow, Frankish et al. 2012) used in that study. Only 30% of gencode transcripts are also included in Refseq annotation. (<http://gencodegenes.wordpress.com/2013/01/08/comparing-different-publicly-available-genesets-against-gencode-7/>). We calculated the allele frequency for each of the variants by dividing the number of alleles (count 1 for heterozygous and 2 for homozygous) with that variant by the total number of allele positions (2 times the total number of people). We first calculated the allele frequency for each ethnic group, and then we took the average over these to obtain the overall average allele frequency for that variant. The density of non-synonymous SNVs and the density for synonymous SNVs for each gene are calculated by dividing the number of corresponding SNVs by the base length of the cDNA sequence. The density of intron

SNVs for a gene is calculated by dividing the number of single base variants in introns of that gene by the total base length of introns of that gene. For each NCBI gene ID, we randomly pick one isoform for the calculation if multiple isoforms are available.

### **Human mRNA expression data**

The Human mRNA expression data for 79 human tissues (Su, Wiltshire et al. 2004) was downloaded from <http://biogps.org/downloads/>. For each gene, the maximum expression level across all tissues was extracted for this study.

### **SNV impact prediction**

We use our SNPs3D profile method (Yue and Moulton 2006) to predict the impact of the non-synonymous variants. We were able to make predictions for 145,336 of the non-synonymous variants, covering roughly 80% of all the high reliability non-synonymous SNVs annotated in this study. The SNPs3D predictor returns a SVM score for each variant. The more negative the score, the more likely that the variant has a high impact. In the study of relationship between SNV impact and SNV frequency, we annotate those SNVs with negative SVM score as predicted deleterious and SNVs with positive SVM score as predicted neutral.

## **Human Protein Functional Interaction Network**

The Human protein functional interaction network (Wu, Feng et al. 2010) was downloaded from <http://genomebiology.com/content/supplementary/gb-2010-11-5-r53-s3.zip> on April 2012. The functional interaction set, FI, is used in this study, and consists of both experimental data and predictions. The network contains both curated functional interactions from biological pathway databases such as Reactome (Matthews, Gopinath et al. 2009) and KEGG (Kanehisa and Goto 2000) and functional interactions based on high throughput experimental data such as protein-protein physical interactions, yeast two hybrid, gene co-expression.

## **Partial correlation**

Pearson partial correlations in this study are calculated using the R package ppcor (Kim and Yi 2006) with variance-covariance matrix inversion option.

## **Quantitative estimation of the selection pressure caused by mRNA expression level and number of functional interactions**

We use a simple linear model to fit the relation between non-synonymous variant density and logarithm of mRNA expression level and logarithm of functional interactions: The formula is:

$$d = K_0 + K_1 \cdot \log_2 E$$

In which  $d$  is the non-synonymous variant density,  $K_0$  and  $K_1$  are constants and  $E$

could be either mRNA expression level or the number of interactors in the functional interaction network.

#### **GO enrichment analysis of outliers**

The GO enrichment analysis was performed on the Gorilla (Eden, Navon et al. 2009) website <http://cbl-gorilla.cs.technion.ac.il/>. The “Two unranked lists of genes” mode was used.

#### 2.4: Discussion

In this study, we addressed a series of questions in evolutionary biology using population genomics data from the 1000 genomes study (Durbin, Altshuler et al. 2010). We first looked at the relationship between the frequency of SNVs and their predicted impact on protein function. We observed that the fraction of predicted deleterious ns-SNVs among all ns-SNVs drops as the population frequency increases. This is as expected from the theory of purifying selection (Akey, Zhang et al. 2002) and similar results have been shown in the 1000 genome exon pilot project (Marth, Yu et al. 2011).

Using population genomics data, we can not only analyze variant distributions in coding regions but can also study intron divergence, providing a better reference set of SNVs under near-zero selection. Based on partial correlation analysis of the relationship between intron SNV density, non-synonymous SNV density, and

synonymous SNV density, we propose that the correlation between synonymous and non-synonymous variant density is the outcome of the combined effect of shared selection pressure at synonymous and non-synonymous sites, and the variation of local mutation prevalence across the genome. We expect that the correlation between dS and dN is largely for the same reasons.

We then look at two factors that are tightly coupled to the protein evolutionary rate: mRNA expression level and number of interactions in the protein network. We found the corresponding population genomics measure of dN, the non-synonymous SNV density, negatively correlates with mRNA expression level. The correlation is stronger for SNVs that have predicted high impact on protein function. The predictor is trained on monogenic disease causing mutations from the HGMD (Stenson, Ball et al. 2003) database. Our previous studies have shown that loss of protein stability is a major cause (approximately 75% of mutations) of monogenic disease. So most high impact predictions are expected to be SNVs with high structural impact, related to misfolding and/or stability loss. This is consistent with the mistranslation-induced-misfolding model proposed by Drummond and Wilke (Drummond and Wilke 2008), which argues the selection pressure on non-synonymous sites is largely caused by the nasty consequences of protein mis-folding. By analyzing the relationship between non-synonymous SNV density and mRNA expression level, we can quantitatively estimate how this biophysical restraint influences the pattern of variant. Using a simple linear model, we estimate the non-synonymous variants density drops 0.18 variants per 1000 nucleotides, equivalent to about 1/25 of the mean value, for each

doubling of expression level.

We did not observe a correlation between synonymous SNV density and the mRNA expression level, which is at variance with the results from inter-species analysis, since a correlation between dS and mRNA expression rate has been reported in (Drummond and Wilke 2008). Three factors may contribute to this difference: 1) The SNV density only counts the number of distinct SNVs in the region of a gene but doesn't consider the population frequency of these variants, which contains important information for the selection pressure of the corresponding sites. In practice, the data are dominated by SNVs with a frequency of less than 0.5%, reducing the selection diversity. 2) We find that the correlation coefficient between non-synonymous variant density and mRNA expression level is about two fold smaller than correlation between dN and mRNA expression level (-0.13 vs -0.2~-0.25 for prefrontal cortex) reported in previous studies (Drummond and Wilke 2008). Since the correlation between dS and mRNA expression level (-0.1) is much weaker than the correlation between dN and mRNA expression level (-0.2~-0.25), the correlation between synonymous variants density and mRNA expression may be too low to be detected. Results from the recent whole genome sequencing of 21 breast cancers (Nik-Zainal, Alexandrov et al. 2012) show that the cancer somatic mutation rate is also correlated with the mRNA expression level. The same pattern has also been found in malignant melanoma and a small cell lung cancer (Pleasant, Cheetham et al. 2010; Pleasant, Stephens et al. 2010). It has been proposed (Nik-Zainal, Alexandrov et al. 2012) that the reason for this observation is that NER (Nucleotide Excision Repair) is recruited

more effectively to highly transcribed genes. This mechanism may not play an important role in the negative correlation we observed from population genomics data. Otherwise, we would also see correlation between synonymous variants and mRNA expression level because the NER doesn't differentiate non-synonymous versus synonymous sites. Further studies are required to understand the difference between the mutation accumulation process in cancer cells and in human populations.

The observed negative correlation between protein-protein functional interactions and non-synonymous SNV density demonstrates that the number of interactions of a protein impose selection on its non-synonymous sites. The selection could be because highly connected genes play a central role in the biological system, so that they are less likely to accept non-synonymous mutations (Wuchty 2004). Another explanation is that a larger proportion of the amino acids in highly connected genes are directly involved in the protein-protein contacts, thus they are subject to constraint caused by their interaction partners (Fraser, Hirsh et al. 2002). We estimate that the non-synonymous SNV density will drop 0.21 per Kb when the number of functional interactions is doubled, and the selection pressure is independent of that caused by mRNA expression level.

The work described here makes use of just 1000 genomes. Many more genomes are becoming available, and these will lead to more accurate calculation of variant density and hence more reliable determination of the factors discussed. For example, a recent exome deep re-sequencing project (Tennessen, Bigham et al. 2012) obtained

exon sequences for 2440 individuals at an average depth of 111X and again found there are large quantities of rare SNVs in the human exome, most of which were previously unknown. Another recent study (Fu, O'Connor et al. 2012) sequenced 6,515 individuals of European American and African American ancestry and inferred the age of 1,146,401 autosomal single nucleotide variants (SNVs). All these upcoming studies will help us know more about the evolutionary process of human SNVs, their origin, history, and the forces that determine their fate. These questions are of great interest for basic research and also important for variant-diseases studies (Do, Kathiresan et al. 2012). We hope with more data becoming available, we can understand the process of human evolution at a new level.

## Chapter 3: GWAS and drug targets

### 3.1: Introduction

Until recently, information on which variants within the human genome contribute to increased risk of common human disease was fragmentary and often statistically weak. New chip-based technologies and large-scale sequencing have now provided relatively unbiased and reliable information on SNVs (single nucleotide variants) and indels significantly associated with altered risk for a number of common diseases. To date, most information has been obtained through genome wide association studies (GWAS) using microarray technology, providing information only on common SNVs (the single nucleotide polymorphisms, SNPs). The current generation of GWA studies typically include several thousand individuals with the disease of interest and a similar number of control individuals without the disease, and in total, more than 1600 loci where variants are associated with complex traits have been discovered (the GWAS catalog, <http://www.genome.gov/gwastudies>)

There have been a number of discussions about the efficacy of GWA studies (Corvin, Craddock et al. 2010). In spite of the success in discovering disease associations, it is becoming clear that many mechanism genes with the highest effect on disease phenotypes are not discovered by GWAS. Studies of blood pressure provide a striking example. There is a long history of identification of genes affecting blood pressure using non-genomic methods, and 30 genes discovered in this way have provided successful targets for treating hypertension (Johnson, Newton-Cheh et al.

2011). But only a few of these candidate genes and no drug targets are discovered in GWAS (Ehret 2010). Limited coverage of current microarrays only explains a fraction of these missing genes (Sober, Org et al. 2009). Further, mouse knockout data suggest that some of the missing genes have very large effect sizes, with blood pressure changes of 10s of mm of Hg (Takahashi and Smithies 2004), whereas the largest changes associated with marker SNPs in GWAS studies are between about 0.5 and 1 mm of Hg (Takahashi and Smithies 2004),

Known drug targets - genes that have a large effect size on the corresponding disease phenotype, and so should be found by GWAS – provide a means of investigating whether non-discovery of mechanism genes is a general phenomenon. In this project, we compared a set of reported mechanism genes in the GWAS catalog ([www.genome.gov/gwastudies](http://www.genome.gov/gwastudies) (Hindorff, Sethupathy et al. 2009), January 2012) with a corresponding set of known drug target genes (obtained from Drugbank (Knox, Law et al. 2011), January 2012) for the same diseases and found the overlap is very low. We also investigated two possible explanations for the low overlap between these two sets of genes and discussed the relationship between these two sets of genes. Finally, we considered the relationship between GWAS genes and drug targets in the context of a protein functional interaction network, and developed a machine learning method to predict new drug targets using the relationship between GWAS genes and known drug targets.

### 3.2: Results

**Comparison of the GWAS catalog and Drugbank shows GWAS only detects a very small fraction of existing drug targets.**

We examined the relationship between genes in the GWAS catalog (Hindorff, Sethupathy et al. 2009) and drug target genes in Drugbank (Knox, Law et al. 2011). The GWAS catalog (<http://www.genome.gov/gwastudies/>) is a comprehensive collection of results from published GWAS studies on a wide variety of disease and other traits such as height. Drugbank (Knox, Law et al.) is a database that combines detailed drug (i.e., chemical, pharmacological and pharmaceutical) data with comprehensive drug target information (sequence, structure, and pathway). We compiled a list of diseases in the GWAS catalog and extracted the reported genes for each of them. We then found the drugs used in treatment of these diseases in Drugbank, and extracted the drug target genes for each drug. Thus, for each disease, we have a list of GWAS reported genes and a list of drug targets. For the 88 GWAS diseases that have drugs in Drugbank, there are on average 29.2 GWAS reported genes and 24.0 drug targets for 19.9 drugs (Table 3.1). In total, only 20 of 856 drug target genes are discovered in GWAS studies of the same disease. This is slightly larger than the estimated overlap of 5 from a completely random model, but still is a very low number if we expect the two sets of genes are related to each other. In Drugbank, some drug targets do not have a known mechanism and are probably “predicted” drug targets based on sequence similarity to the verified drug targets (Imming, Sinning et al. 2006; Overington, Al-Lazikani et al. 2006). We therefore

compiled a list of verified drug targets all of which have known drug action mechanisms documented in Drugbank. We find similar results to those for the complete list of drug targets. For those 353 drug targets for 81 diseases with known mechanisms and with corresponding GWAS studies, only 12 are discovered by GWAS (Table 3.2). On average, there are 30 GWAS reported genes and 11.2 verified drug targets for each of these 81 diseases.

Table 3.1 Overlapping between GWAS reported genes and drug targets

Disease	Number of Drugs	GWAS reported genes	Number of drug targets	GWAS overlap, same disease	GWAS overlap, all diseases
Acute lymphoblastic leukemia	6	19	10	0	3
Age-related macular degeneration	9	23	2	1	2
Allergic rhinitis	69	11	20	0	5
Alzheimer's disease	5	54	179	0	40
Amyotrophic lateral sclerosis	3	26	2	0	1
Ankylosing spondylitis	39	17	29	0	9
Arthritis	168	7	112	0	35
Asthma	102	43	52	1	19
Atopic dermatitis	12	8	3	0	1
Atrial fibrillation	45	7	25	0	14
Attention deficit hyperactivity disorder	3	81	1	0	1
Autism	3	6	10	0	5
Basal cell carcinoma	6	8	9	0	2
Bipolar disorder/ Schizophrenia	93	215	110	1	32

Blood pressure/ Hypertension	351	100	114	3	35
Breast cancer	84	42	43	1	13
Celiac disease	3	74	1	0	0
Chronic kidney disease	8	69	6	0	2
Chronic lymphocytic leukemia	14	17	29	0	5
Chronic myeloid leukemia	6	9	15	0	6
Chronic obstructive pulmonary disease	14	18	7	0	2
Colorectal cancer	8	14	16	0	6
Coronary heart disease	6	84	5	0	3
Crohn's disease	7	136	23	0	9
Cystic fibrosis	8	7	11	0	5
Depression/ Depressive disorder	45	68	73	0	17
Diabetes	46	205	59	4	21
Duodenal ulcer	8	2	18	0	5
Emphysema	10	5	17	0	5
Endometrial cancer	1	2	2	0	0
Endometriosis	5	4	7	0	3
End-stage renal disease	2	2	8	0	3
Epilepsy	18	1	53	0	10
Esophageal cancer	1	18	2	0	1
Gallstones	1	1	1	0	0
Gastric cancer	2	3	1	0	0
Glaucoma	24	13	31	0	6
Glioblastoma	2	1	1	0	0
Heart failure	51	16	65	0	27
HIV/AIDS	54	62	53	1	9
Hodgkin's Lymphoma	8	7	31	0	7
Hypertriglyceridemia	2	5	4	0	3
Hypothyroidism	5	43	8	1	5
Inflammatory bowel disease	2	18	8	0	4

Kawasaki disease	1	20	11	1	5
Malaria	17	3	17	0	4
Male infertility	6	5	3	0	3
Melanoma	9	20	6	0	0
Menopause	9	23	15	0	4
Migraine	20	7	46	0	10
Multiple myeloma	7	3	10	0	3
Multiple sclerosis	10	123	30	1	12
Myocardial infarction	29	14	44	0	17
Narcolepsy	2	4	6	0	1
Nephropathy/ Nephrotic syndrome	20	26	38	0	9
Neuroblastoma	2	2	6	0	2
Non-small cell lung cancer	5	7	10	0	1
Obesity	4	40	11	0	4
Osteoarthritis	26	3	46	0	10
Osteoporosis	13	10	10	0	2
Ovarian cancer	5	10	4	0	1
Paget's disease	4	9	6	0	1
Pancreatic cancer	2	29	11	0	4
Panic disorder	6	10	18	0	4
Parkinson's disease	20	62	184	1	34
Polycystic ovary syndrome	2	7	2	0	1
Prostate cancer	14	94	21	0	8
Psoriasis/Psoriatic arthritis	19	30	39	0	13
Refractive error	1	4	4	0	1
Restless legs syndrome	2	6	18	0	6
Rheumatoid arthritis	46	67	80	2	29
Sleepiness	1	2	2	0	0
Stevens-Johnson syndrome/ toxic epidermal necrolysis	1	12	1	0	0
Stroke	8	4	7	0	6
Tardive dyskinesia	3	1	22	0	7
Testicular cancer	4	7	6	0	2

Thyroid cancer	2	5	3	0	2
Tuberculosis	12	5	18	0	4
Type 1 diabetes	8	74	18	0	8
Type 2 diabetes	28	91	34	3	13
Ulcerative colitis	5	95	9	1	6
Uterine fibroids	1	7	1	0	0
Venous thromboembolism	1	7	3	0	2
Vitiligo	4	25	8	1	2
Mean	19.90	29.18	24.00	0.26	7.09

**Analysis using 1000 genomes project data shows Drug Target genes have fewer high frequency deleterious non-synonymous SNPs than GWAS reported genes**

We next consider why GWAS identifies so few known drug targets. A study of all the SNPs in the GWAS catalog (Hindorff, Sethupathy et al. 2009) shows that reported SNPs are common (median risk allele frequency 36%, interquartile range (IQR) 21%-53%), and are associated with modest effect size (median odds ratio 1.33, IQR 1.20-1.61 ).

We speculated that drug target genes escape GWAS studies because they contain few common SNPs with negative effect. To test this hypothesis, we look at the distribution of SNP frequencies and SNP effect size in GWAS identified genes and drug targets. SNP frequencies were calculated from 1000 genomes project (Durbin, Altshuler et al. 2010) data.

It's not easy to measure the effects of SNPs, considering these can come from many

different mechanisms (altering protein sequence, altering the regulation of the expression of genes, changing the splicing pattern, changing the stability of messenger RNA). We focus our investigation on non-synonymous SNPs for two reasons: 1) A study has shown that reported SNPs are significantly overrepresented in non-synonymous sites (Hindorff, Sethupathy et al. 2009). 2) There are a number of computational methods to estimate the effect of non-synonymous SNPs on *in vivo* protein function, with useful accuracy (Ng and Henikoff 2003; Yue, Melamud et al. 2006; Bromberg and Rost 2007; Adzhubei, Schmidt et al. 2010). We calculated the frequency of all the non-synonymous SNPs in GWAS identified genes and drug targets, and then predicted the effect of these SNPs using two methods, the SNPs3D profile method (Yue and Moulton 2006) and PolyPhen2 (Adzhubei, Schmidt et al. 2010). We found the drug targets genes do have fewer non-synonymous SNPs (0.0155/aa. vs. 0.0171/aa.) and the tendency is more significant for common (Allele frequency > 5%) non-synonymous SNPs (0.00169/aa. vs. 0.00221/aa, Mann-Whitney test  $P=0.0017$ ). For ns-SNPs with a predicted negative effect, there is also significant difference between these two set of genes. The fraction of predicted negative effect ns-SNPs among all common ns-SNPs is smaller for drug targets than GWAS reported genes (15.8% vs. 19.2%). A possible explanation for the low occurrence of common high impact SNPs is that the activity level of drug targets genes is strongly coupled to the disease phenotype. As a result they are under relatively high selection pressure, and SNPs with a substantial impact on function will be eliminated or tend to be at a low frequency. We expect this rule applies to both non-synonymous SNPs and other mechanism SNPs, although we could only test it on non-synonymous ones.

Table 3.3 Comparison of non-synonymous SNP distribution between GWAS reported genes and drug targets

	Drug Targets	GWAS reported genes	All genes
Density of Non-synonymous SNPs	0.0155	0.0171	0.0171
Density of Common Non-synonymous SNPs	0.00169 P=0.0017 <sup>1</sup> P=0.0023 <sup>2</sup>	0.00221	0.00214
Fraction of Common deleterious Non-synonymous SNPs predicted to be deleterious	15.8% (SNPs3D)  18.4% (PolyPhen2)	19.2% (SNPs3D)  22.3% (PolyPhen2)	22.4% (SNPs3D)  23.5% (polyPhen2)

<sup>1</sup>P-value for Mann-Whitney test against the density of common non-synonymous SNPs for GWAS reported genes. <sup>2</sup>P-value for Mann-Whitney test against the density of common non-synonymous SNPs for all genes.

**Evolutionary analysis shows drug target genes are under slightly stronger negative selection than GWAS reported genes**

If the drug targets genes are under stronger selection as we proposed, we should observe it through evolutionary analysis. dN/dS (Kimura 1977) is a measure of selection pressure on genes. We compared the dN/dS for these two set of genes using data from H-invDB (Yamasaki, Murakami et al. 2008) and found both are under

stronger selection (Table 3.4) than all genes. We also looked at the selection effect on monogenic disease genes from the HGMD database (Stenson, Ball et al. 2003) which are considered to be under negative selection because mutations in those genes cause disease. We found HGMD genes are also under negative selection in recent history (dN/dS calculated using human-chimp orthologs). The selection against variants in drug target genes is slightly stronger than that against variants in GWAS reported genes (Table 3.4) for dN/dS calculated using Human-chimp orthologs, suggesting the selection is stronger for drug targets in recent history.

Table 3.4 dN/dS analysis for GWAS reported genes and drug targets

		Number of data points	Mean dN/dS	P Value for Mann-Whitney test against all genes	P Value for Mann-Whitney test against gwas reported genes
Human-Mouse orthologs	All genes	13691	0.22		
	GWAS reported genes	2932	0.19	2.44e-09*	
	Drug targets	1035	0.18	1.21e-04*	0.43
	Drug targets with known mechanism	432	0.17	6.04e-06*	0.038*
Human- Chimpanzee orthologs	All genes	14173	0.44		
	GWAS reported genes	2911	0.36	1.26e-13*	
	Drug targets	1020	0.33	2.78e-13*	0.0098*
	Drug targets with known mechanism	423	0.32	4.2e-8*	0.013*

### **The influence of transcript length**

We also examined an additional possible reason why GWAS does not identify known drug targets. Suppose the mechanism SNPs are distributed randomly across the relevant genes, then the chance of a gene being identified should be related to its length.

A disease by disease analysis shows that for most of the diseases the GWAS reported genes are significantly longer than the drug target genes (paired Mann-Whitney test,  $P=1.89e-6$ ) and in general GWAS reported genes tend to be longer than drug target genes and longer than all other genes (Figure 3.1). The mean longest transcript length for GWAS reported genes is about 110K while the mean longest transcript length for drug targets is about 60K.

Thus, these two factors, a tendency for GWAS genes to be longer than drug targets, and a lower density of common SNPs in drug targets at least partially explain the observed low overlap between these two set of genes.

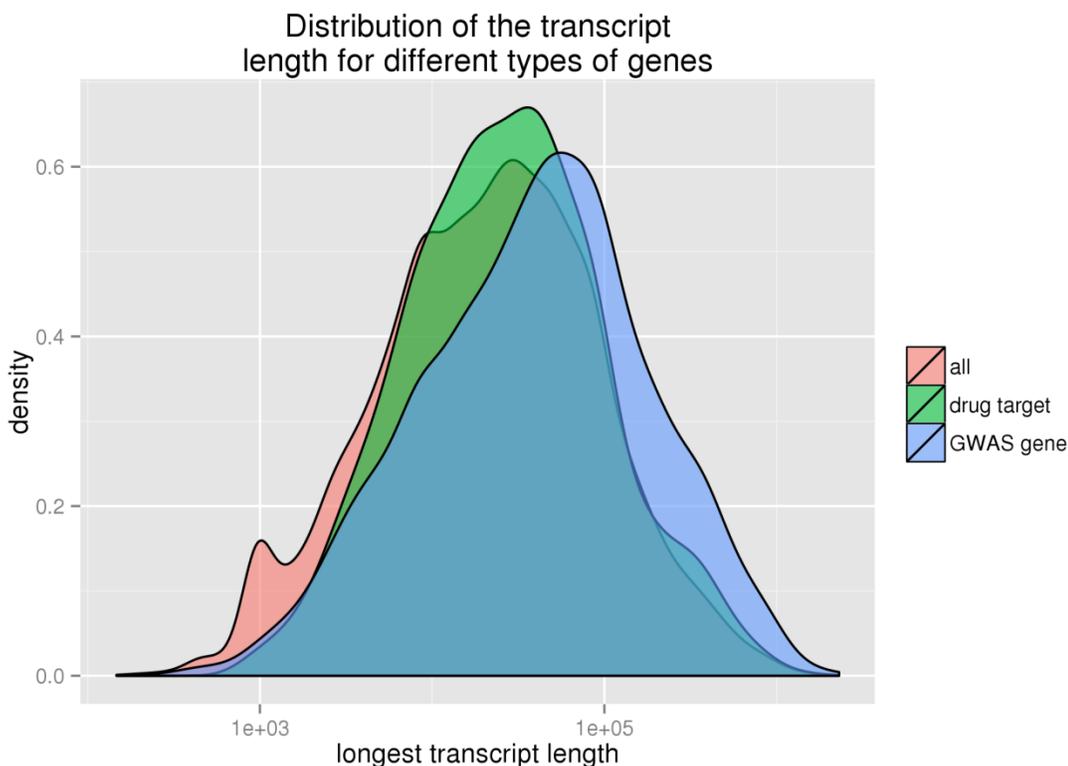


Figure 3.1 Distribution of the log longest transcript length for different types of genes.

**Network analysis shows GWAS reported genes are close to drug target genes in a biological network**

Although most drug targets are not identified through GWAS studies, they are obviously as much involved in the disease mechanism as GWAS genes, and so will have similar properties, particularly in terms of pathway and network relationships. A number of studies have incorporated network information to aid in identifying various classes of genes, for example using a network module formalism to combine signals from multiple GWAS studies (Jia and Zhao 2011; Jia, Wang et al. 2012) and using

network flow models to predict drug targets from expression and other data in prostate cancer (Yeh, Yeh et al. 2012). Network models have also been used to identify pathways implicated in cancer (Ciriello, Cerami et al. 2012). It has already been observed that GWAS genes are substantially more closely connected in a functional network (Rossin, Lage et al. 2011) than random genes, and we expect that to be the case for other large effect genes, such as known drug targets.

There are many resources available for different types of human biological networks. Protein-Protein interaction data (Chatr-aryamontri, Ceol et al. 2007; Breitkreutz, Stark et al. 2008) have a wide coverage but usually have a high false positive rate. Curated pathways like KEGG (Kanehisa 2002; Kanehisa, Goto et al. 2002; Kanehisa, Goto et al. 2004) and BioCarta (<http://www.biocarta.com/genes/index.asp>) are considered to be accurate but the coverage is sparse (Wu, Feng et al. 2010). Networks built from other kinds of relationship such as regulatory networks deduced from micro-array data (Lee, Hsu et al. 2004; Prieto, Risueno et al. 2008) or networks based on biochemical reactions (Lang, Stelzer et al.) are too narrow in terms of the interactions they capture.

In this study, we use the Functional Interaction (FI) network from (Wu, Feng et al. 2010), which is a protein functional interaction network generated by extending curated biological pathways with non-curated sources of information, including protein-protein interactions, gene co-expression, protein domain interaction, Gene Ontology (GO) annotations and text-mined protein interactions, and covers about

50% of the human genes. The network strikes a balance between experimentally validated results and prediction, with the prediction part benchmarked by a reasonably rigorous process. We were able to map 611 out of 821 drug targets genes and 1125 out of 1914 GWAS reported genes for the 88 diseases to the network.

Examination of the network proximity of GWAS genes to each other and to drug targets for the same disease indeed shows a strikingly close-knit matrix of relationships. Figure 3.2 shows the network formed for the 43 GWAS and 16 drug target genes (Knox, Law et al. 2011) for Type I Diabetes that project onto the FI network, and only including genes from these two sets which are linked by not more than one other intermediate gene. All drug targets and all but five of the GWAS genes form part of a single continuous sub-network.

This suggests that the two sets of genes are relatively close in their biological function. One way to look at the relationship between GWAS reported genes and drug target genes is to measure how close each GWAS gene is to its nearest drug target (Figure 3.3). Different drug targets may be involved in different mechanisms, so that a GWAS gene may be very close to one drug target while further from others. The distributions show that the distances from a GWAS reported gene to the closest drug target are on average much shorter than those of a random gene to a closest drug target, and the shortest distance from a drug target gene to the closest GWAS reported gene is also shorter than that of a random gene to the closest GWAS reported gene. Notably, drug targets are about three fold enriched in the neighbors of GWAS genes

and are also enriched in GWAS second neighbors (genes two steps away in the gene network) (Figure 3.3).

Highly connected genes have more neighbors, and thus are more likely to include GWAS genes as neighbors. Thus, the observed enrichment of short paths between drug targets and GWAS genes could partially be a consequence of higher connectivity for drug targets. To control for this effect, we compared the degrees of drug targets with all genes (Figure 3.3C), and we found drug targets have a slightly higher level of degree (Mann-Whitney test,  $P = 0.014$ ) on average. However the difference is marginal, and is unlikely to significantly contribute to the substantial difference between the short path distribution for drug targets and all genes.

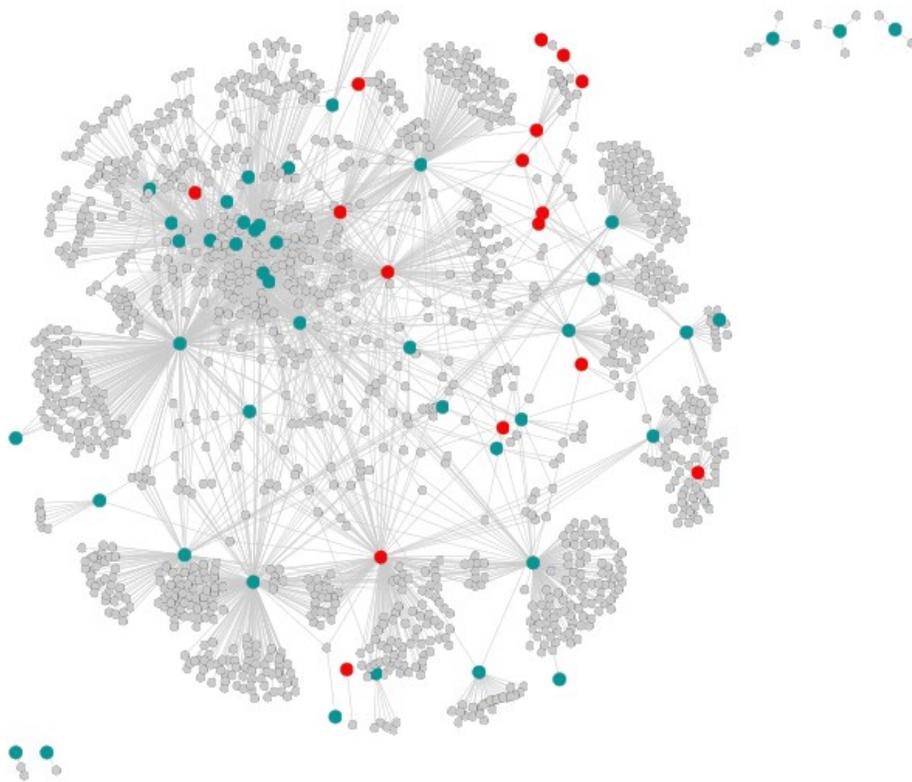


Figure 3.2 Network for GWAS gene and drug targets of Type I Diabetes  
Continuous network substructure formed by 43 of the 74 GWAS (**green**) and 16 of the 18 drug targets (**red**) for Type 1 Diabetes, allowing not more than one intermediate gene (**grey**). GWAS and drug target genes are intermingled in the network, and short paths are sufficient to form a connected network for almost all genes. FI network, figure from Cytoscape.

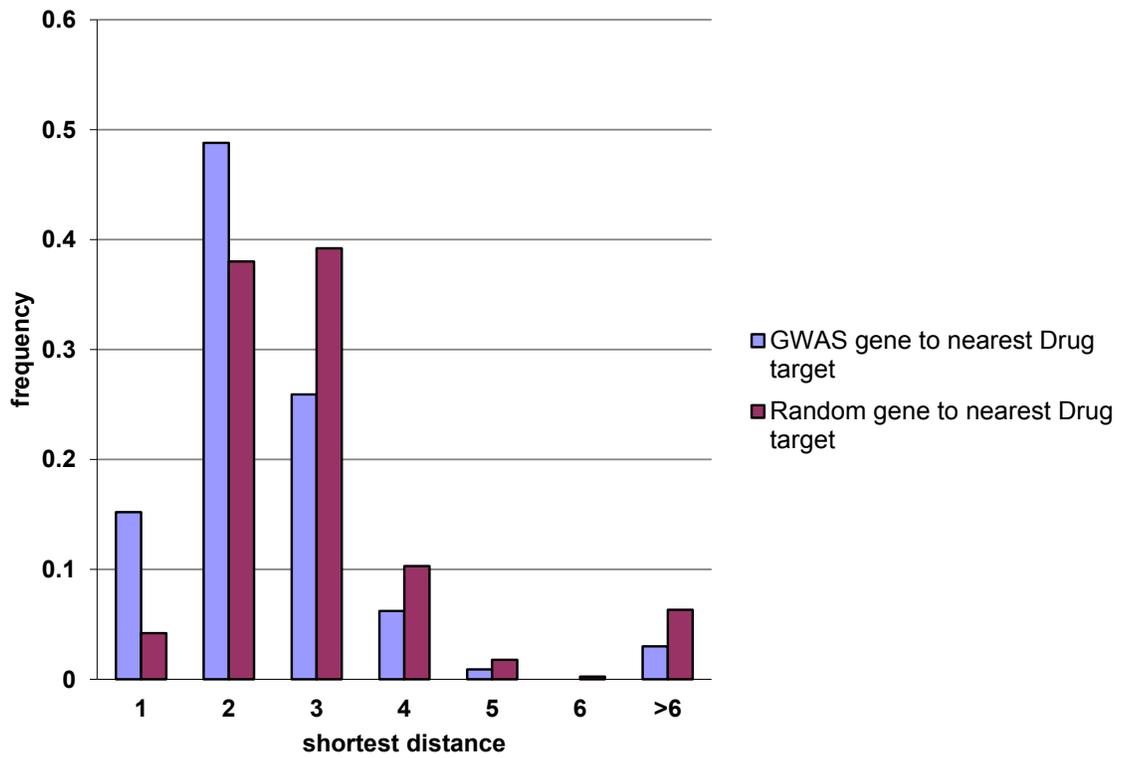


Figure 3.3A The distribution of shortest distance to the nearest drug target for GWAS reported genes and all genes.

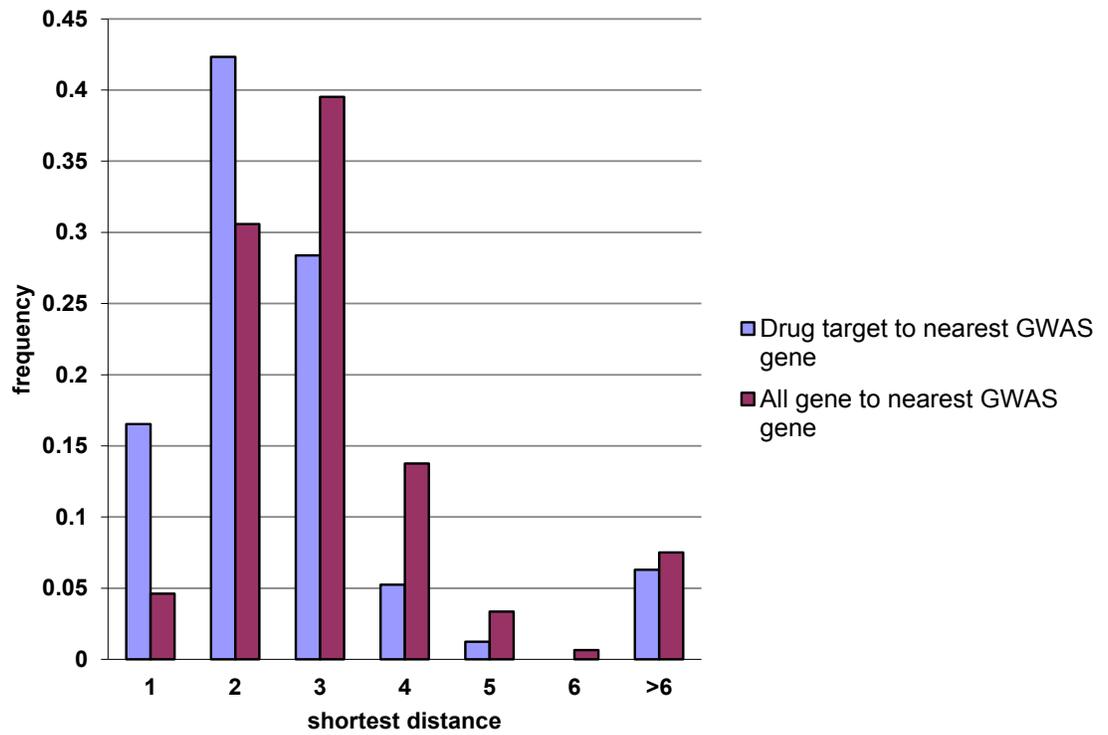


Figure 3.3B The distribution of the shortest distance to the nearest GWAS genes for drug targets and all genes.

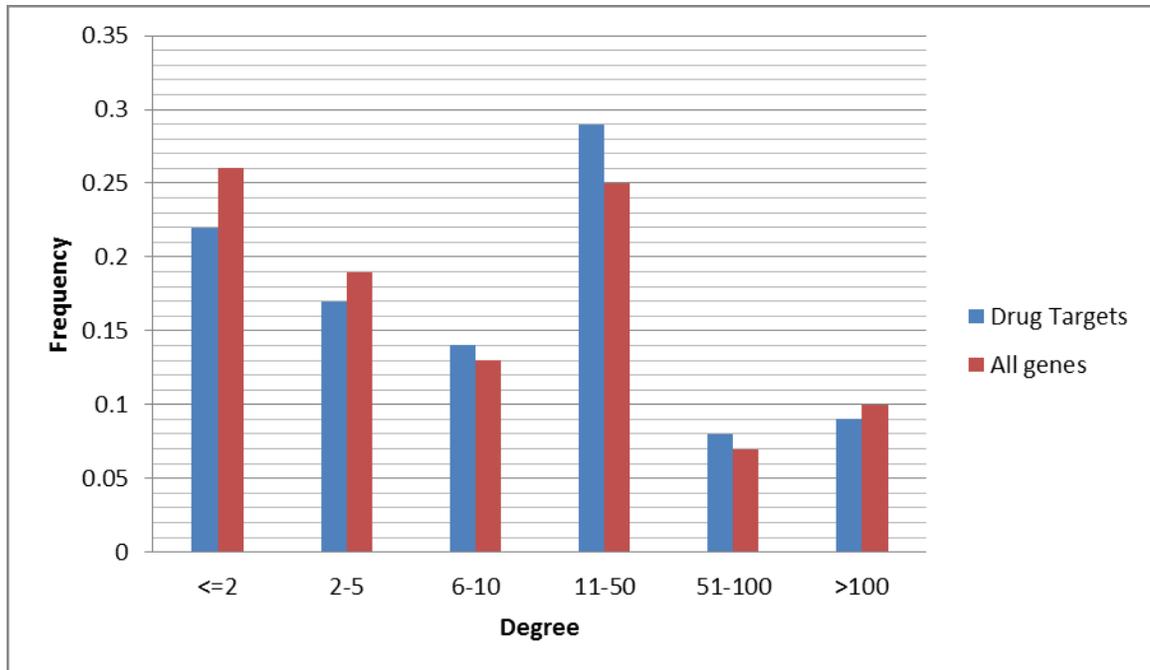


Figure 3.3C Distribution of degree for drug targets and all genes in the FI network.

The drug targets have a slightly higher level of degree (Mann-Whitely test  $P = 0.014$ ).

The difference is marginal, and does not explain the observed enrichment of drug targets in the first neighbor of GWAS reported genes.

### Machine learning method for drug target discovery

The strong relationship between drug targets and GWAS gene revealed in the network analysis led us to think about ways to identify drug targets from GWAS genes using machine learning methods based on network features. The idea is to evaluate the probability that any gene is a potential drug target, given its network environment. Since we observed a threefold enrichment of drug targets in the first neighbors of the GWAS genes, we use the number of GWAS neighbors for a gene as

a feature. This quantity is highly dependent on the total number of neighbors a gene has, so we also use the degree of the gene as a control. As the previous analysis shows, second neighbors of drug targets genes (genes that are two steps away in the protein interaction network) are also enriched for GWAS genes, thus we also use the number of GWAS genes in the second neighbors of the genes as a feature. These three features capture the enrichment information from our previous analysis, but there are some subtle relationships not included. The problem of identifying drug targets is very similar to the problem of finding missing relationships in social network analysis. We therefore also use common friends with GWAS genes, a widely used feature in the social network machine learning field (Fire, Tenenboim et al. 2011). The common neighbor feature is defined as the proportion of neighbors shared by two genes:

$$\text{Common Neighbor}(A,B) = \frac{\text{count}(N_A \cap N_B)}{\text{count}(N_A \cup N_B)}$$

In which  $N_A$  is the set of Neighbors for gene A,  $N_B$  is the set of Neighbors for gene B.

The total number of features for each gene is  $3+N$ , where  $N$  is the number of GWAS genes for that disease that are mapped to the protein network. Since the number of drug targets (average 30) for a disease is very small compared to the total number of genes in the FI network (10956), the training set is highly unbalanced if we use the latter as the true negative set. To address this issue, we focus on the 932 existing drug targets in Drugbank that are also in the FI network, and thus restrict the task to

identifying existing drugs that can potentially be repurposed to treat other diseases. Repurposing is an attractive goal, since such use is much easier than developing a new drug from scratch (Carley 2005).

We include the 30 diseases with at least 10 approved drug targets and 10 GWAS genes in the FI network. We tested four machine learning methods using the WEKA software package (Witten, Frank et al. 1999): a SVM with a polynomial kernel, a SVM with RBF kernel, a Naïve Bayes Network, and Random Forests. Among these we found the best result is achieved by a Random forest (Table 3.5). The best case is Kawasaki disease, with a true positive rate of 70% (recovering seven out of 10 known drug targets) and a false positive rate of 2.7%.

Table 3.5 Machine learning result for different diseases, using a Random Forest.

Disease	GWAS genes	Drug targets (Mapped in Network)	True Positive	False Positive	Precision	Recall	ROC area	F-Measure
Ankylosing spondylitis	17	29(24)	0.36	0.123	0.074	0.36	0.73	0.123
Menopause	24	15(14)	0.571	0.098	0.082	0.571	0.819	0.143
Multiple sclerosis	126	30(28)	0.393	0.052	0.19	0.393	0.75	0.256
Myocardial infarction	14	44(40)	0.175	0.135	0.055	0.175	0.571	0.084
Nephropathy/Nephrotic syndrome	26	38(35)	0.371	0.245	0.056	0.371	0.576	0.097
Obesity	40	11(11)	0.273	0.098	0.032	0.273	0.724	0.058
Osteoporosis	10	10(10)	0.2	0.189	0.011	0.2	0.546	0.022

Pancreatic cancer	29	11(6)	0.167	0.1	0.011	0.167	0.611	0.02
Panic disorder	10	18(16)	0.438	0.118	0.061	0.438	0.754	0.107
Parkinson's disease	62	184(132)	0.606	0.226	0.307	0.606	0.712	0.407
Asthma	43	52(47)	0.213	0.102	0.1	0.213	0.713	0.136
Prostate cancer	95	21(18)	0.5	0.073	0.118	0.5	0.686	0.191
Psoriasis/Psoriatic arthritis	31	39(36)	0.5	0.076	0.209	0.5	0.852	0.295
Rheumatoid arthritis	67	80(68)	0.324	0.131	0.163	0.324	0.677	0.217
Type 1 diabetes	76	18(16)	0.25	0.104	0.04	0.25	0.631	0.07
Type 2 diabetes	92	34(28)	0.214	0.116	0.054	0.214	0.595	0.086
Bipolar disorder/Schizophr enia	217	110(81)	0.593	0.15	0.273	0.593	0.744	0.374
Blood pressure/Hypertens ion	101	114(102)	0.412	0.143	0.261	0.412	0.717	0.319
Breast cancer	43	43(38)	0.289	0.072	0.147	0.289	0.745	0.195
Chronic lymphocytic leukemia	17	29(26)	0.423	0.098	0.11	0.423	0.653	0.175
Colorectal cancer	14	16(16)	0.25	0.154	0.028	0.25	0.53	0.05
Acute lymphoblastic leukemia	19	10(10)	0.7	0.07	0.097	0.7	0.889	0.171
Crohn's disease	139	23(22)	0.455	0.093	0.105	0.455	0.764	0.171
Depression/Depres sive disorder	68	73(62)	0.597	0.172	0.198	0.597	0.722	0.297
Diabetes	209	59(51)	0.216	0.081	0.134	0.216	0.712	0.165
Allergic rhinitis	11	20(19)	0.263	0.128	0.041	0.263	0.589	0.071
Glaucoma	14	31(25)	0.16	0.189	0.023	0.16	0.443	0.04
Alzheimer's disease	54	179(125)	0.544	0.178	0.321	0.544	0.69	0.404
Heart failure	16	65(54)	0.481	0.222	0.118	0.481	0.655	0.189
HIV/AIDS	63	53(34)	0.353	0.121	0.099	0.353	0.715	0.155

Kawasaki disease	20	11(10)	0.7	0.027	0.219	0.7	0.919	0.333
------------------	----	--------	-----	-------	-------	-----	-------	-------

---

**Potential new drug targets for drug repurposing**

The “false positive” drug targets are those drug targets for other diseases which have very similar network properties to those for the disease under study. These may indeed be mistakes made by the classifier. However, a more optimistic view would be that some of these “false positive” drug targets are good candidates for repurposing, not discovered before.

For example, at the top of the false positive list for the best case, Kawasaki disease, we found C1QB and C1QC, both subcomponents of complement C1Q. C1Q has been shown to be associated with lupus erythematosus (Bowness, Davies et al. 1994; Korb and Ahearn 1997; Walport, Davies et al. 1998), which is another autoimmune disease closely related to Kawasaki disease (Laxer, Cameron et al. 1988; Diniz, Almeida et al. 2012). C1Q is the target of several FDA approved drugs, for example, Etanercept, a drug treating rheumatoid arthritis and Adalimumab, a drug treating rheumatoid arthritis, psoriatic arthritis, ankylosing spondylitis, and other immune system mediated diseases. Thus these drugs may be potential candidates for use against Kawasaki disease.

Another disease where the method performs well is acute lymphoblastic leukemia, with a false positive rate of 7% and a true positive rate of 70%. We have a relatively

long list of “false positive” targets (Table 3.6). Careful inspection of these genes reveals some that may have relevance to acute lymphoblastic leukemia, and so drugs for which they are targets provide potential candidates for repositioning. For example, chromosomal aberrations (chromosome translocation) involving FGFR1 are associated with stem cell myeloproliferative disorder and stem cell leukemia lymphoma syndrome (provided by RefSeq, Jul 2008). FGFR1 is the drug target of Palifermin, which is a recombinant human keratinocyte growth factor (KGF) for the treatment of oral mucositis associated with chemotherapy and radiation therapy. It’s also the drug target for several experimental drugs.

Previous studies found differential expression of the oncogene RET in acute myeloid leukemia (Gattei, Degan et al. 1998), a distinct but related leukemia. In the version of Drugbank used in this analysis, there is no drug targeting RET for the treatment of acute lymphoblastic leukemia. Recently, however, the drug Ponatinib has been approved by FDA for treatment of Philadelphia chromosome positive acute lymphoblastic leukemia (Ph+ALL) that is resistant or intolerant to prior tyrosine kinase inhibitor therapy. Thus, one of the suggested drug target has been confirmed, although for use with a new drug.

Table 3.6 Top “false positive” drug targets for acute lymphoblastic leukemia.

Target	Description from NCBI	Random Forest Probability
--------	-----------------------	------------------------------

---

MAPK3	The protein encoded by this gene is a member of the MAP kinase family. MAP kinases, also known as extracellular signal-regulated kinases (ERKs), act in a signaling cascade that regulates various cellular processes such as proliferation, differentiation, and cell cycle progression in response to a variety of extracellular signals.	1
PIK3R1	Phosphatidylinositol 3-kinase plays an important role in the metabolic actions of insulin, and a mutation in this gene has been associated with insulin resistance.	0.96
RAF1	v-raf-1 murine leukemia viral oncogene homolog 1	0.96
EGFR	Mutations in this gene are associated with lung cancer. Multiple alternatively spliced transcript variants that encode different protein isoforms have been found for this gene	0.96
FGFR2	Mutations in this gene are associated with Crouzon syndrome, Pfeiffer syndrome, Craniosynostosis, Apert syndrome, Jackson-Weiss syndrome, Beare-Stevenson cutis gyrata syndrome, Saethre-Chotzen syndrome, and syndromic craniosynostosis.	0.96
KDR	This receptor, known as kinase insert domain receptor, is a type III receptor tyrosine kinase. Mutations of this gene are implicated in infantile capillary hemangiomas.	0.94
FLT1	This protein binds to VEGFR-A, VEGFR-B and placental growth factor and plays an important role in angiogenesis and vasculogenesis.	0.94
FGFR1	Chromosomal aberrations involving this gene are associated with stem cell myeloproliferative disorder and stem cell leukemia lymphoma syndrome.	0.94
IL2RG	The protein encoded by this gene is an important signaling component of many interleukin receptors	0.92
ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog	0.92
FGFR3	This particular family member binds acidic and basic fibroblast growth hormone and plays a role in bone development and maintenance. Mutations in this gene lead to craniosynostosis and multiple types of skeletal dysplasia.	0.9
AKT1	v-akt murine thymoma viral oncogene homolog 1	0.9
INSR	insulin receptor	0.9

---

---

IL2RA	Mutations in this gene are associated with interleukin 2 receptor alpha deficiency.	0.9
SDC2	The syndecan-2 protein functions as an integral membrane protein and participates in cell proliferation, cell migration and cell-matrix interactions via its receptor for extracellular matrix proteins. Altered syndecan-2 expression has been detected in several different tumor types.	0.88
MAPK1	The protein encoded by this gene is a member of the MAP kinase family. MAP kinases, also known as extracellular signal-regulated kinases (ERKs), act as an integration point for multiple biochemical signals, and are involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation and development.	0.86
CD247	The protein encoded by this gene is T-cell receptor zeta, which together with T-cell receptor alpha/beta and gamma/delta heterodimers, and with CD3-gamma, -delta and -epsilon, forms the T-cell receptor-CD3 complex.	0.86
RET	ret proto-oncogene	0.86
VEGFA	vascular endothelial growth factor A	0.86
PTPN1	protein tyrosine phosphatase, non-receptor type 1	0.86
IL3RA	The protein encoded by this gene is an interleukin 3 specific subunit of a heterodimeric cytokine receptor.	0.84
HDAC1	histone deacetylase 1, Together with metastasis-associated protein-2, it deacetylates p53 and modulates its effect on cell growth and apoptosis.	0.82
CCND1	The protein encoded by this gene belongs to the highly conserved cyclin family, whose members are characterized by a dramatic periodicity in protein abundance throughout the cell cycle. This protein has been shown to interact with tumor suppressor protein Rb and the expression of this gene is regulated positively by Rb. Mutations, amplification and overexpression of this gene, which alters cell cycle progression, are observed frequently in a variety of tumors and may contribute to tumorigenesis	0.82
FASN	fatty acid synthase	0.82
CD4	The protein functions to initiate or augment the early phase of T-cell activation, and may function as an important mediator of indirect neuronal damage in infectious and immune-mediated diseases of the central nervous system.	0.8

---

### 3.3: Methods

#### **Connecting GWAS reported genes with drug targets through the drug indication information from Drugbank.**

GWAS reported genes: We downloaded the GWAS catalog from <http://www.genome.gov/admin/gwascatalog.txt> in Jan 2012 and manually filtered out non-disease traits and combined studies for each disease. ‘Reported genes’ were extracted to provide the list of GWAS genes for each disease.

Drug targets: We downloaded complete Drugbank data from <http://www.drugbank.ca/downloads> in January 2012. For each of the diseases in our GWAS gene list, we find corresponding drugs for that disease by searching the “indication” information for all drugs in Drugbank. Then for each of these drugs, we extract all of the corresponding target genes.

Verified drug targets: We pick drug targets with the entry “Pharmacological action” labeled as “Yes” in the Drugbank.

GWAS reported genes and drug targets were mapped to NCBI gene IDs to provide unique identifiers for comparison. The full set contains 1914 GWAS reported genes and 821 drug targets for 88 diseases. It contains a total of 4013 GWAS reported genes and 1463 drug target genes if we drop the restriction for the 88 diseases. The verified drug target set has 353 genes for 81 diseases. For each disease, we compare the list of GWAS reported genes and drug targets and find the overlap between these two lists.

### **Calculating expected overlap between GWAS reported genes and drug targets using a complete random model**

We estimate there are 20,000 human genes. For a specific disease, there are 'm' GWAS reported genes, and there are 'n' drug targets for this disease. Thus the expected random overlap between the two gene lists for that disease is  $n*m/20000$ . We calculated the expected overlap for each disease and then added them up to get the expected total number of overlap between drug targets and GWAS reported genes for the same disease.

### **SNP impact analysis for GWAS genes and drug target genes**

1000 genome VCF data are downloaded from <http://www.1000genomes.org/data>. The 2010 November data set is used. We extracted all non-synonymous variants from 1000 genome data based on the Refseq annotation downloaded from UCSC genome browser on Jan 2012, and calculated the allele frequency for each of the non-reference variants by dividing the number of alleles (count 1 for heterozygous and 2 for homozygous) by the number of total possible (2 times the number of people).

We found non-synonymous SNPs in the coding regions of 3550 out of the 4013 GWAS reported genes and 1249 out of the 1463 drug targets. We used our SNPs3D profile method (Yue, Melamud et al. 2006) and PolyPhen2 (Adzhubei, Schmidt et al. 2010) to predict the impact of these non-synonymous SNPs on protein function. We obtained SNPs3D profile predictions of impact for SNPs in 1041 out of 1249 drug

targets and 2764 out of 3550 GWAS reported genes. We also obtained PolyPhen2 predictions for 1227 out of 1249 drug target genes and 3484 out of 3550 GWAS reported genes.

The density of non-synonymous SNPs in each gene is calculated by dividing the number of non-synonymous SNPs for that gene by the length of that gene's protein sequence provided by the UCSC genome browser <http://genome.ucsc.edu/>. One splicing form is randomly chosen for each NCBI gene ID.

The fraction of common negative impact SNPs are calculated as follows: For the SNPs3D profile method, the fraction is calculated by dividing the total number of common negative impact SNPs (allele frequency > 5% and SNPs3D score < 0 ) in a group of genes by the total number of common SNPs having a prediction score from SNPs3D in that group of genes. For PolyPhen2, the fraction is calculated by dividing the total number of common negative impact SNPs (allele frequency > 5% and Polyphen2 prediction “possibly damaging” or “probably damaging”) in a group by the total number of common SNPs having Polyphen2 predictions in that group.

### **Transcript length analysis**

The longest transcript for each drug target and GWAS reported genes was picked based on the Refseq annotation downloaded from UCSC genome browser on Jan 2012.

### **Evolutionary analysis for GWAS reported genes and Drug target genes**

Ratios of non-synonymous to synonymous substitution rates, dN/dS, for human proteins were downloaded from <http://www.h-invitational.jp/evola/download.html> in March 2012. The h-inv (Imanishi and Nakaoka 2009) IDs were converted to NCBI Gene IDs using a conversion map downloaded from <http://biodb.jp/download.cgi>. dN/dS from Human-Mouse orthologs and Human-Chimpanzee orthologs were selected. Human-Mouse dN/dS are considered to reflect selection over a relatively long time period, and Human-Chimpanzee dN/dS to reflect more recent history.

### **Human gene network analysis for GWAS reported genes and drug target genes**

The Functional Interaction protein network (Wu, Feng et al. 2010) was downloaded from <http://genomebiology.com/content/supplementary/gb-2010-11-5-r53-s3.zip>. This un-weighted map consists of 209,988 functional interactions involving 10956 proteins, and covers roughly half of the human coding genome. Gene symbols in this data set were converted to NCBI gene IDs. Finally 1125 out of 1914 GWAS reported genes and 611 out of 821 drug target genes for the 88 diseases and 932 drug targets of all 1463 drug targets were mapped into the network.

The Floyd-Warshall algorithm (Floyd 1962) was used to calculate the shortest path between all gene pairs in the network. The resulting set of inter-node distances serves as a background distribution. For each disease, we extracted the set of all pairwise

distances between GWAS genes for that disease, between drug targets genes, and between GWAS genes and drug target genes. For each disease, we also calculated the shortest path from every gene in the network to the nearest GWAS gene for that disease and to the nearest drug target for the disease.

### **Machine learning for drug targets**

We use a random forest implemented in WEKA (Witten, Frank et al. 1999) to train on  $N+3$  features to predict known drug targets for a disease from all drug targets. The training set is unbalanced because the number of drug targets for each disease is very small (median 28) compared to all possible drug targets, 932. We use the MetaCost procedure (Domingos 1999) to deal with the unbalanced training set, which gives more penalty to false negative errors than to false positive errors. We set the cost factor to be the ratio between the number of “correct” and “incorrect” drug targets. We set the parameter  $K$ , the number of separating features as the square root of the number of all features and set the parameter  $I$ , the number of decision trees in the random forest, as 50. 10 fold cross validation was used to measure the performance for the random forest method for each disease.

### 3.4: Discussion

In this chapter, we began by evaluating the capability of GWA studies to identify existing drug targets. From an analysis using Drugbank and the GWAS catalog, we found that these studies find a surprisingly small fraction of drug targets.

Studies (Hindorff, Sethupathy et al. 2009) have shown that GWAS methods typically find high frequency SNPs with modest phenotype effects. On the other hand drug targets have big effect sizes with respect to disease phenotypes. We expect fewer high frequency deleterious SNPs in these genes and do observe this trend for non-synonymous SNPs through analysis of population genomics data from the 1000 genome project.

Although we only look at the impact and distribution pattern of non-synonymous SNPs in these two set of genes, it's likely that SNPs exerting their influence through other mechanisms (for example, altering the regulation of the expression of genes, changing the splicing pattern, changing the stability of messenger RNA) also follow the same pattern because the selection pressure is the same for all kinds of impact mechanisms. Through analysis of dN/dS for these two set of genes, we again found the selection pressure acting on drug targets is stronger than for GWAS reported genes.

If drug targets genes are under strong negative selection so that they harbor fewer common SNPs with high negative effect, why do GWA studies not detect SNPs in these genes that have mild negative effects? To answer this question, it's important to understand the spectrum of effect size generated by SNPs. For non-synonymous mutations, an analysis of the impact of mutations on the activity for seven different enzymes (Yampolsky and Stoltzfus 2005) shows that the majority of negative effect

mutations are highly deleterious, leading to at least a 10 fold decrease of assayed activity. So mild-effect non-synonymous SNPs may be rare because of an intrinsic property of protein mutations.

The results are disappointing if our aim is discovering new drug targets directly from GWAS. However the close relationship between drug targets and GWAS reported genes makes the GWAS genes valuable reference points for finding new drug targets. Based on network features of human protein networks for GWAS reported genes and known drug targets, we developed a machine learning method to predict potential drug targets suitable for a specific disease from all existing drug targets in Drugbank. One of our initial proposals for new drug targets has now been confirmed.

The present GWAS relies on common SNPs. With the development of new sequencing technology, high quality exome sequencing will to replace the role of DNA chips in GWAS studies (Kiezun, Garimella et al. 2012), and we already begin to gain more insight into the role of rare variants. For example, a recent exome deep re-sequencing project (Tennesen, Bigham et al. 2012) found there are a large number of rare SNVs in the human exome, most of which were previously unknown. A deep re-sequencing project for drug target genes has also found an abundance of rare functional variants (Nelson, Wegmann et al. 2012). These rare variants in drug target genes may play a role in risk of complex disease. All these data provide both opportunity and challenge for computational biology. Because of difference in distribution pattern and effect size between common SNPs and rare SNVs, new

statistical procedures need to be designed to detect association between rare SNVs and human diseases (Do, Kathiresan et al. 2012). For example, pooled association tests (Price, Kryukov et al. 2010) which combine rare variants in the same gene together provide a promising approach to address the problem. Another more straightforward way to establish the relationship between genes and diseases is to simply find genes that are mutated with high frequency in the diseases, a practice adopted in many cancer studies (Kumar, White et al. 2011; Wang, Kan et al. 2011; Wei, Walia et al. 2011). For diseases like hypertension, many candidate genes have been discovered using non-genomic methods (Johnson, Newton-Cheh et al. 2011). Rare variants in these candidate genes in patients will be of great interest and may reveal key mechanisms in the development of the corresponding diseases as well as provide clues for potential therapeutics.

## Chapter 4: Improved missense variant modeling accuracy using an Alignment Quality Estimator

### 4.1: Introduction

Missense single nucleotide variants in the human population are one of the principal causes of Mendelian disease (Stenson, Ball et al. 2009), and a major contributor to predisposition for cancer (King, Marks et al. 2003) as well as somatic driver mutations in these diseases (Shi and Moulton 2011). They also play a significant role in common complex trait diseases (Gorlatova, Chao et al. 2011). More than three dozen computational methods have been developed to estimate the *in vivo* functional impact of these substitutions.

Most of these methods rely on the phylogenetic pattern of residue use at the substitution position, utilizing the tendency of uncommon residues to have deleterious effect.

For example, SIFT (Ng and Henikoff 2003) and the SNPs3D profile module (Yue, Melamud et al. 2006) predict the effect of a missense substitution based on the conservation of that amino-acid in close related species and the dissimilarity between the original amino acid and the substituted one. The PolyPhen2 (Adzhubei, Schmidt et al. 2010) method employs a Bayesian model trained on a hybrid set of features consisting of structural properties as well as sequence profiles. Although sequence

based methods in principle capture all influences on fitness, they have the disadvantage of providing no direct insight into the underlying molecular mechanism. In particular, they are unable to distinguish between effects on protein function, such as altered enzymatic activity, allosteric regulation, or interaction with binding partners, and effects on protein folding and stability. In earlier work, we showed that about 75% of monogenic disease causing mutations act through these latter mechanisms. The SNPs3D structural module (Yue, Melamud et al. 2006) is a purely structure based prediction method aimed at identifying effects on protein thermodynamic stability. It uses an atomic level model of the protein to calculate 15 physical and chemical features from the interactions of the original and substituted amino acids. A support vector machine (SVM), trained on mutations causative of monogenic disease and a control set of interspecies variants, is then used to predict whether or not stability is significantly affected (typically a change of 2 Kcal/mol or greater).

A limitation of the SNPs3D and other structural approaches is the need for an adequately accurate three-dimensional structure. At present, experimental structures cover only about 8% of known missense variants. Comparative modeling (Eswar, John et al. 2003) based on a template of a related structure provides a means of extending this coverage, but the more remote the template from the structure of interest, the less accurate the model. Previous benchmarking of the SNPs3D method showed that down to 40% sequence identity between the template and the target structure, there is little decline in accuracy (Yue and Moult 2006). Inclusion of these

models doubles the coverage of missense variants to 16%. If it were possible to extend coverage down to models based on 25% sequence identity, coverage would increase to 38%. Thus, we seek strategies for better coverage.

There are two possible ways to increase coverage further than 40%. First, build more accurate models. A great deal of effort is already being placed into improving modeling methods, and although progress is being made, this is a tough problem (Kryshtafovych, Fidelis et al. 2011). Here we take the second approach, which is to identify those models built with existing methods that are already sufficiently reliable for the specific application of estimating the effect of a substitution on stability.

The primary cause of the deterioration of model quality as template sequence identity falls is errors in alignment of the target and template sequences (Kryshtafovych, Fidelis et al. 2011). Misalignment by one residue position in the sequence causes a backbone error of 3.8Å, large enough for the values of the 15 features used in the SVM to be very seriously distorted. We have therefore developed a method that estimates the reliability of the alignment at the position of the amino acid substitution. The method employs the Smith-Waterman dynamic programming procedure to find the alignment between the target and template sequences with the maximum score (Smith and Waterman 1981; Waterman 1983). When the sequence identity between the target and template is low, it's possible to generate several alternative alignments with marginal score differences. When this happens, it's hard to tell which of these alignments is the best. If we simply pick the one with the highest score, chances are

that we are picking a wrong alignment. In order to pick high confidence alignments, we developed an algorithm called Force-Bypass. The method perturbs the initial best alignment at the amino-acid position of interest to generate a new alignment differing at that position. The alignment score difference between the new alignment and the original one (score drop) is converted to the probability that the original alignment is correct at that position, using a calibration from a structurally determined set of alignments.

We applied this algorithm to pick variants at high confidence alignment positions in low sequence identity models for variants in the SNPs3D training data. Retraining and benchmarking delivered significantly higher prediction accuracy on these compared to the low confidence ones. We then applied the method to pick high confidence low sequence identity models for missense SNVs from 1000 genomes project (Durbin, Altshuler et al. 2010), and by doing this, increased the number of model-able missense SNVs from 16% to 23%. The Force-bypass method developed in this paper can also be used in various situations of evolutionary analysis, in which the sequence identity is very low but the requirement for the accuracy at specific positions is stringent.

#### 4.2: Results

##### **Score Drop is a good criterion for Alignment quality**

As noted above, the Force-Bypass procedure is implemented in the widely used

Smith-Waterman (Smith and Waterman 1981; Waterman 1983) protein sequence alignment algorithm. The goal of this algorithm is to maximize the total alignment score, which is the sum of the residue pair scores over all positions.

Residue pair scores may be calculated in a number of ways. We use a simple substitution matrix, and profile-profile matching. The substitution matrix reflects how “similar” a pair of aligned amino acids is. There are many different versions of substitution score matrices. Here we use the popular BLOSUM62 matrix (Henikoff and Henikoff 1992). In this matrix the substitution score between Lysine and Arginine is 2, reflecting the fact that they are highly similar amino acids, while the substitution score between Tryptophan and Arginine is -3, reflecting the fact that they are very different and pairing between them should be avoided in the alignment.

Given pairwise scores for all possible matches of each residue in one sequence with each residue in the other, the Smith-Waterman algorithm employs dynamic programming to calculate the alignment with the highest score. That alignment is a best estimate of the historical equivalence of positions in the sequences. When the identity between the two sequences is low, several alignments may have similar scores. The ranking of these alignments depends heavily on the underlying pairwise scoring procedure, thus the alignment with the highest score is not necessarily the correct one. The similarity of scores for alternative alignments with respect to a particular residue is thus related to the probability that alignment is correct.

To determine how robust the alignment at a particular position is we perturb the dynamic programming algorithm to generate an alternative alignment which is the

one with the highest score of all alignments with an alternative pairing at the position we are interested in. Then we compare the alignment score of the alternative alignment with the highest score one (details in Methods). The difference between the scores of the two alignments is used as a measure of the reliability of the alignment at that position. To use this procedure to assign accuracy of alignments, we must have a gold standard alignment set to train our algorithm. Balibase (Thompson, Koehl et al. 2005) is a set of manually curated multiple alignments of low identity protein sequences based on protein structural superimposition. The database is widely used for benchmarking sequence alignment algorithms (Karplus and Hu 2001). We convert the multiple sequence alignments in Balibase into pair-wise alignments. We then aligned each pair of sequences in Balibase with the standard Smith-Waterman algorithm. After that, we compared the pair-wise alignments extracted from Balibase with the alignments generated by the Smith-Waterman algorithm. We sorted the positions in the Smith-Waterman alignments into two categories: Correct ones, where Smith-Waterman is consistent with its Balibase counterpart; and incorrect ones, where Smith-Waterman is inconsistent with its Balibase counterpart. We then used our perturbation procedure to bypass each position in turn and calculate the score drop.

Figure 4.1A shows the comparison of the score drop for these two categories of positions using the BLOSUM62 matrix. Correct alignments have a very distinct score distribution from the incorrect ones.

### **Comparison between profile-profile alignment and substitution matrix based alignment**

Matching residue pairs on the basis of a profile-profile comparison produces more accurate alignments than those obtained using a single substitution matrix (von Ohlsen, Sommer et al. 2003; Ohlson, Wallner et al. 2004), and has been especially popular in the field of structural modeling. Profile-profile alignment is often performed with the Smith-Waterman algorithm or equivalent dynamic programming procedures (Rychlewski, Jaroszewski et al. 2000; Yona and Levitt 2002; von Ohlsen, Sommer et al. 2003). The profile reflects the possibility of observing each type of amino-acid at each position in the sequence, and is usually generated by PSI-BLAST (Altschul, Madden et al. 1997). Then a score function is designed to provide a score between any pair of profiles aligned at a position. We implemented the weighted-sum (von Ohlsen, Sommer et al. 2003) score function and performed the same procedure on the Balibase set as with the BLOSUM62 substitution matrix. We found that the profile-profile methods are more likely to get a correct alignment (Figure 4.1B). I.e., the proportion of consistent positions is higher (86%) in the profile-profile alignments than in the substitution matrix based alignments (76%). But the coverage is lower than that using the substitution matrix (Table 4.1). I.e., the number of aligned positions is smaller.

Table 4.1 Comparison of coverage and accuracy between BLOSUM62 and profile-profile based alignments

	Total positions covered	Consistent with BaliBase	% consistent with Balibase
BLOSUM62 alignments	327923	249669	76%
Profile-Profile alignments	198464	170542	86%

**Probability of a correct alignment as a function of score drop**

We bin the positions according to the score difference, and in each bin, we calculate the fraction of consistent positions as an empirical confidence measure (Figure 4.2A, 4.3A). Using logistic regression, we can fit a confidence value function to each level of score difference (Figure 4.2B, 4.3B).

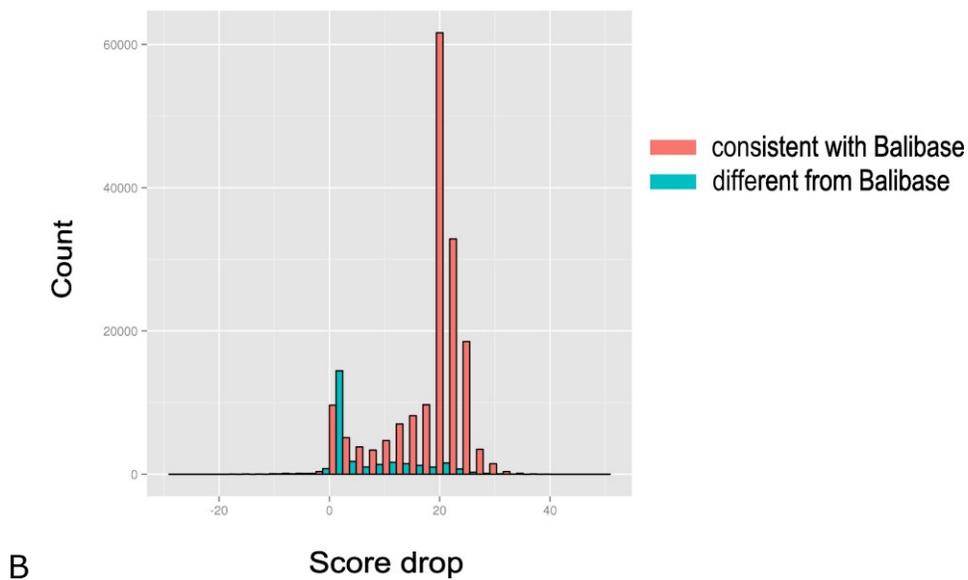
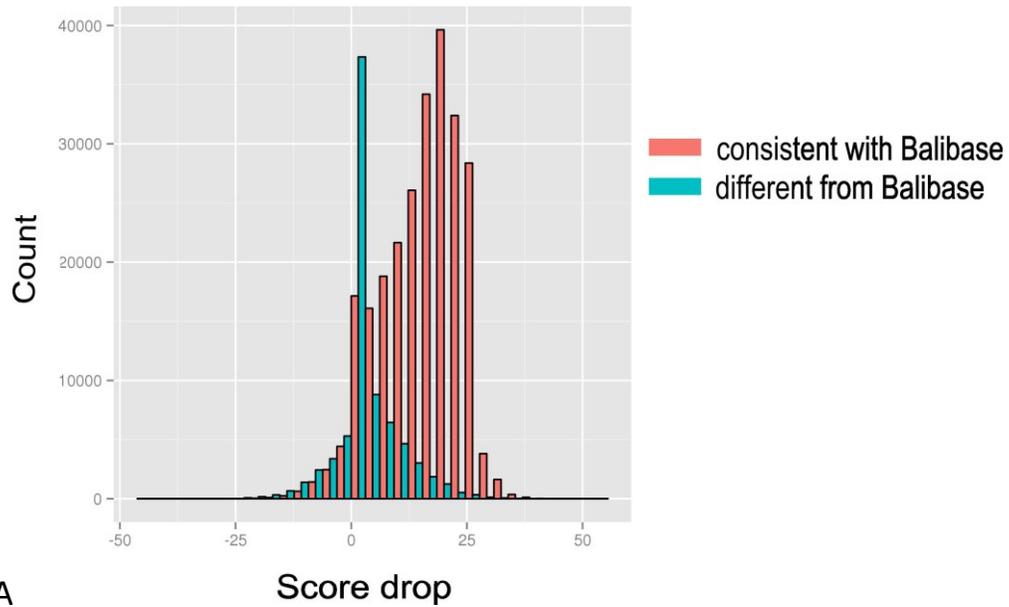
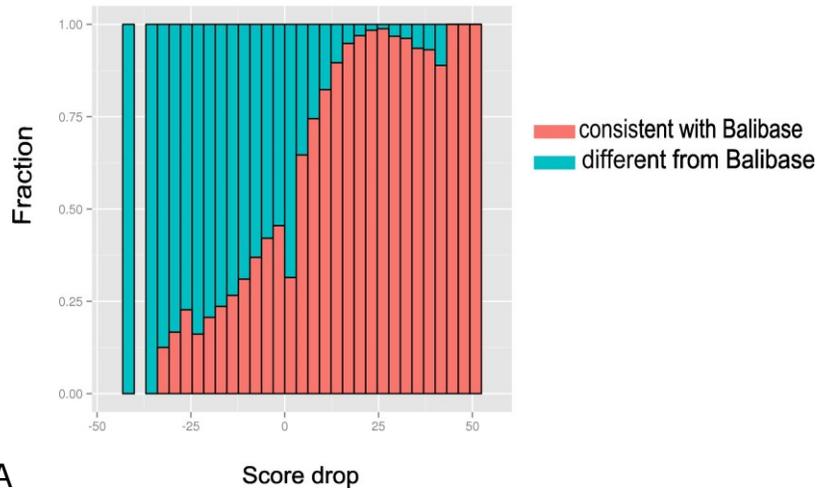
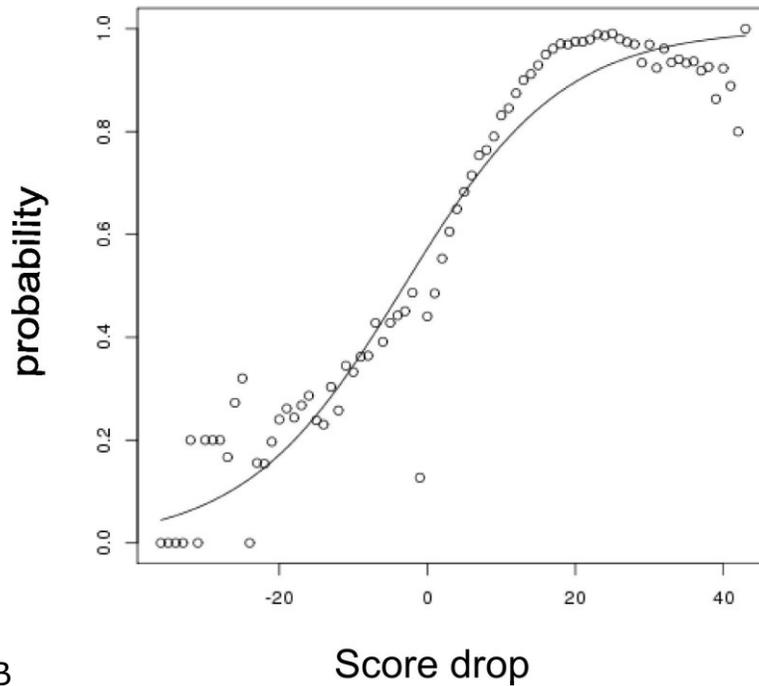


Figure 4.1 Distribution of score drops for positions consistent and inconsistent with Balibase alignments. A) Result generated using the Smith-Waterman algorithm with the BLUSOM62 substitution matrix; B) Result generated using profile-profile scoring.



A



B

Figure 4.2 A) Fraction of positions consistent with Balibase for each score drop value for BLOSUM62 alignment. B) Logistic regression for the probability of a correct alignment based on score drop values.

The fitted curve is  $y = 1 / (1 + \exp(-0.29153 - 0.09338 * x))$

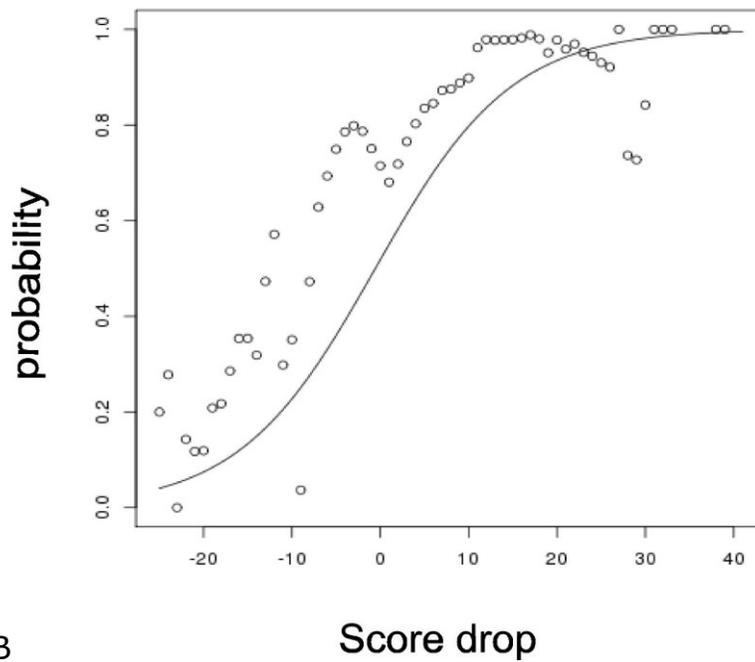
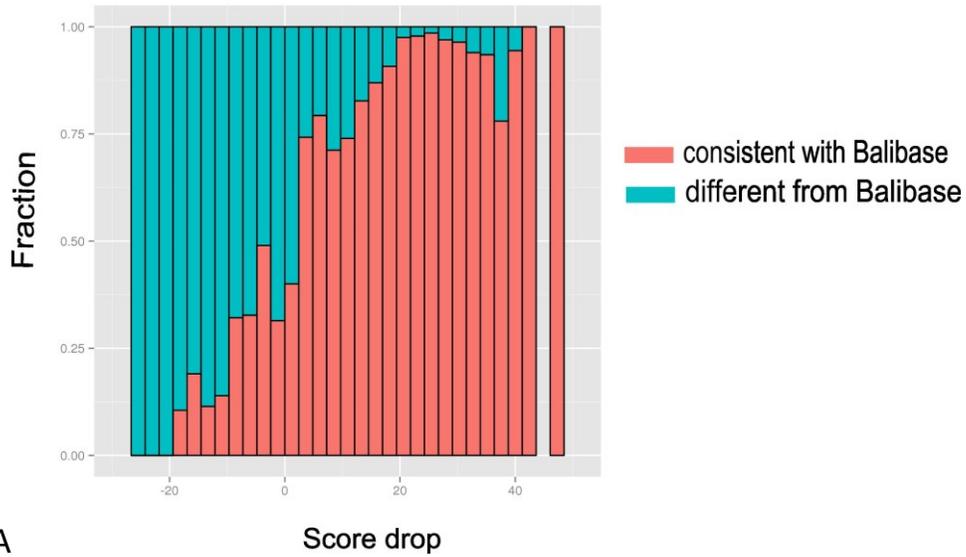


Figure 4.3 A) Fraction of positions consistent with Balibase for each score drop value for profile-profile scoring. B) Logistic regression for the probability of a correct alignment based on score drop values.

The fitted curve is:  $y = 1 / (1 + \exp(-0.07337 - 0.12965 * x))$ .

### **SNPs3D performs better at high confidence alignment positions than at the low confidence ones**

As mentioned in the Introduction, our goal is to improve the quality of missense SNV impact prediction by filtering out low confidence alignments. We trained our SNPs3D structure predictor separately on high confidence alignment positions and on low confidence alignment positions, using low sequence identity models. The details of the SVM training is described in the previous paper (Yue, Melamud et al. 2006). The original predictor is trained on models built from PDB templates with at least 40% sequence identity. Where possible, we replaced each 40%+ sequence identity PDB template with a related structure with a sequence identity to the target protein between 25-40%, and rebuilt the models (see Methods). We then used the Force-Bypass algorithm to generate alternative alignments and calculated score differences for each variant position in these low sequence identity models. We used 10 as a score difference cutoff for both BLOSUM62 and profile-profile alignments, corresponding to a confidence value (the fraction of alignments that are correct) of 0.77 and 0.79 respectively. For each alignment method, we divided the models into two categories: high confidence ones (score drop  $\geq 10$ ) and low confidence ones (score drop  $< 10$ ). The accuracy of predictions is higher for the high confidence models (Table 4.2), although the accuracy of these high confidence models is not as high as for 40%+ identity models (Table 4.2). Note that high false negative rates are expected in these results: The positive training sets contain disease mutations that

affect *in vivo* protein function in multiple ways. We previously estimated (Yue, Li et al. 2005) that about 75% affect thermodynamic stability, and these are the true positives in our training set. On that basis, 25% false negatives is low as should be expected. Nevertheless, it is notable that the low quality models return larger false positives rates than the high quality ones, while the false positive rates are comparable.

Table 4.2 Comparison of SNPs3D performance for different model sets

	False Positive	False Negative
40%-100% identity models	18%	26%
25-40% BLOSUM62 (high confidence)	20.8%	33.8%
25-40% BLOSUM62 (low confidence)	27.1%	43.2%
25-40% profile-profile (high confidence)	21.2%	31.4%
25-40% profile-profile (low confidence)	31.6%	35.3%

### **Examples of low sequence identity, high quality models**

#### **NAGLU F410S**

NAGLU, alpha N-acetylglucosaminidase, is an enzyme that degrades heparan sulfate by hydrolysis of terminal N-acetyl-D-glucosamine residues in N-acetyl-alpha-D-glucosaminides. Defects of enzyme are the cause of mucopolysaccharidosis type IIIB (MPS-IIIB), also known as Sanfilippo syndrome B. This disease is characterized by the lysosomal accumulation and urinary excretion of heparan sulfate (provided by RefSeq, Jul 2008). The mutation F410S is one of the disease mutations according to HGMD (Stenson, Ball et al. 2003). There is currently no template in the PDB database that has 40% percent identity covering this position. The best model we can build uses PDB code 2VCC as a template. This protein is a glycoside hydrolase from *Clostridium Perfringensi*. The alternative alignment has a local score drop of 34.5, far in excess of our threshold of 10, and with an estimated probability of correct alignment of 0.97 (Figure 4.4). The SNPs3D predictor assigns this mutation as a destabilizing, and gives the main mechanism as loss of hydrophobic interactions (Figure 4.4).

**A**

Part of the Alignment between NAGLU and 2vcc\_A

NAGLU	229	VLDQMRSGMTFVLPFAFAGHVEAVTRVFFQVNVTKMGSGHFNCSYSCS	278
2vcc_A	369	MHDRMQSGFINPVLQGTSGNVPDFKFKMQEAGTISQGGN---CGFDRP	414
NAGLU	279	FL LAP----EDPIFPIIGSLFLRELKEFG-TDHIYGADTFNEMQPPSS	322
2vcc_A	415	DMLKTYVNEGEADYFQKVADVFYEQKEVFGDVTNFGVDVDFHE-----	458
NAGLU	323	EPSYLAATA-----VYE---AMTAVDTEAVMLLQGNLFOGPOQFWG	362
2vcc_A	459	-----GGNTGDLNGKIYEIIQNMIEHDNDVAVVIQNW-----QG	494
NAGLU	363	PAQIRAVLGAVPGRLLVLDLFAESQPVYTRTASFOGQPFVCHLRFNFGG	412
2vcc_A	495	NPSNKLGLTKDQMVLDLFSVSPDWRLEE-RDLFWIMMLRFNFGG	543
NAGLU	413	NHGLFGALEAVNGPEAARLFPNST--MVGTQMAPEGISQNVVYSLMAEL	461
2vcc_A	544	NHGMDAAPEKL--ATEIPKALANSEHWGIGITPEAINVNPFLAYELFPM	591
NAGLU	462	GNKDPVPLAAWVTSFAARRYGVSHPDAGAWRLLSRVYNSGCAEGRG	511
2vcc_A	592	AMTRDQI-NFRWTEDYIERRYGKTNKELEAWNIIIDTAYKRNDDYQG	640

Part of the alternative Alignment between NAGLU and 2vcc\_A

NAGLU	229	VLDQMRSGMTFVLPFAFAGHVEAVTRVFFQVNVTKMGSGHFNCSYSCS	278
2vcc_A	369	MHDRMQSGFINPVLQGTSGNVPDFKFKMQEAGTISQGGN---CGFDRP	414
NAGLU	279	FL LAP----EDPIFPIIGSLFLRELKEFG-TDHIYGADTFNEMQPPSS	322
2vcc_A	415	DMLKTYVNEGEADYFQKVADVFYEQKEVFGDVTNFGVDVDFHE-----	458
NAGLU	323	EPSYLAATA-----VYE---AMTAVDTEAVMLLQGNLFOGPOQFWG	362
2vcc_A	459	-----GGNTGDLNGKIYEIIQNMIEHDNDVAVVIQNW-----G	494
NAGLU	363	PAQIRAVLGAVPGRLLVLDLFAESQPVYTRTASFOGQPFVCHLRFNFGG	412
2vcc_A	495	NPSNKLGLTKDQMVLDLFSVSPDWRLEE-RDLFWIMMLRFNFGG	540
NAGLU	413	NHGLFGALEAVNGPEA-----RLFNST--MVGTQMAPEGISQNVVYS	456
2vcc_A	541	---FGGRMGDAAPEKLATEIPKALANSEHWGIGITPEAINVNPFLAYE	586
NAGLU	457	LMAELGNKDPVPLAAWVTSFAARRYGVSHPDAGAWRLLSRVYNSG	506
2vcc_A	587	LLFDMAWRDQI-NFRWTEDYIERRYGKTNKELEAWNIIIDTAYKRN	635

NAGLU	406	MLHNFGGNHG	415
2vcc_A	537	MLHNFGGRMG	546
NAGLU	406	MLHNFGGNHG	415
2vcc_A	537	MLHN-----	540



Local Alignment score drops 34.5

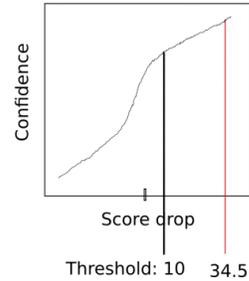
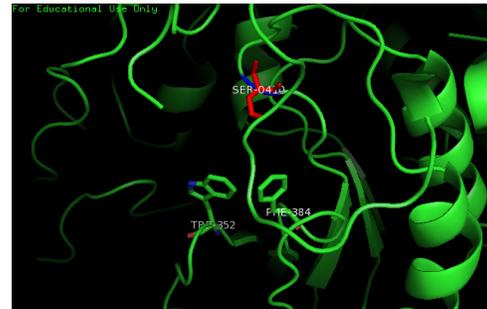
**B****C**

Figure 4.4 (A) Comparison of original and force-bypass alignments between NAGLU and the 2VCC\_A template around the position of the F410S mutation. The alternative alignment doesn't allow F410 in NAGLU to match with F541 in 2VCC\_A, leading to a major gap and an alignment score drops 34.5, well above the

threshold 10, suggesting the original alignment is highly accurate. (B) Structural model for the NAGLU build from 2VCC. From the model we can see the substitution of F410 to a Serine (S) leads to the loss of hydrophobic interaction between F410 to TRP353 and PHE384, resulting in loss of thermodynamic stability.

#### ITGB4 C34Y

ITGB4 is the  $\beta$ 4 subunit of integrin. Integrin mediates cell-matrix or cell-cell adhesion, and transduced signals that regulate gene expression and cell growth. This gene encodes the integrin  $\beta$ 4 subunit, a receptor for the laminins. This subunit tends to associate with the  $\alpha$ 6 subunit and is likely to play a pivotal role in the biology of invasive carcinoma. Mutations in this gene are associated with epidermolysis bullosa and with pyloric atresia. (provided by RefSeq, Jul 2008). The best model we can build from a PDB structural template that covers this disease causing mutation C34Y (Stenson, Ball et al. 2003) has sequence identity 35.9% with the target. That is 3IJE\_B, a crystal structure of the  $\beta$ 3 subunit of integrin. The score drop for the alternative alignment is 28, again well above the threshold 10, and corresponding to probability of alignment correctness of 0.95 (Figure 4.5). The SNPs3D method predicts this mutation to be destabilizing through breakage of a disulfide bridge and overpacking of residues (Figure 4.5).

**A**

part of alignment between 1TGB4 and 3ije\_B

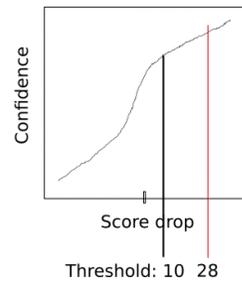
ITGB4	1	NRCKKAPVKSCTECVVRVDKDCAYCTDEM--FRDRRCNTQAE	LAAGCORE	48
3ije_B	3	NICTTRGVSSCQCCLAVSPMCWCSDEALPLGSPRCDLKEN	LLKDNCAPE	52
ITGB4	49	SIVVMESFQITEETQID----TLRRSQMSPOGLRVRLRPGEERHFL		93
3ije_B	53	SIEFFVSEARVLEDRPLSDKSGDSSQVTVQVSPQRIALRLRFPDGSNFSI		102
ITGB4	94	EVEFPLESPVDLYILMDFSNMSDDLNLKMKQNLARVLSQLTSDYITIG		143
3ije_B	103	QVRQVEDYFVDIYLLMDSLYSMKDLSIQNLGKTLATQMKRLTSLNRIG		152
ITGB4	144	PGKPVKVSVPQDMR-PEKLEKPNPNSD---PPFSFKNVLSEEDVDE		188
3ije_B	153	PGAFVDPKVSFMYIISPEALENFCYDMKTTCLPMFGYKHVLTLDQVTR		202
ITGB4	189	FRNKQGERISGNLDAPEGGFDAIQAVCTRDIGWRPDSHLLVFSTES		238
3ije_B	203	FNEEVKQSVSRNDAPEGGFDAIQAVCTDEKIGWRNDASHLLVPTDA		252
ITGB4	239	AFHYEADGANVLGIMSRNDRCHLDTGTQYQYRQDYPSPVTLVRLLA		288
3ije_B	253	KTHIALDGR--LAGIVQPNQDQCHVGSNDHYSASTMDFYSLGLMTEKLS		300

part of alternative alignment between 1TGB4 and 3ije\_B

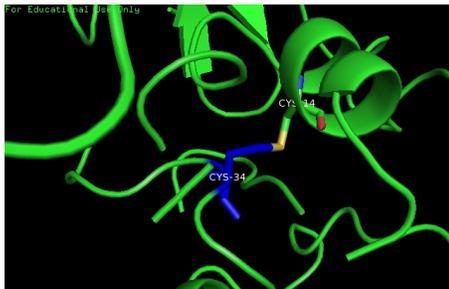
ITGB4	1	NRCKKAPVKSCTECVVRVDKDCAYCTDEMFR----DRRCNTQAE	LLAAG	44
3ije_B	3	NICTTRGVSSCQCCLAVSPMCWCSDEALPLGSPRCDLKEN	LLKDN	48
ITGB4	45	CORESIVVMESFQITEETQIDT----TLRRSQMSPOGLRVRLRPGEER		89
3ije_B	49	CAPESTIEFFVSEARVLEDRPLSDKSGDSSQVTVQVSPQRIALRLRFPDGSK		98
ITGB4	90	HFELEVFPELESVDLYILMDFSNMSDDLNLKMKQNLARVLSQLTSD		139
3ije_B	99	NFSIQVRQVEDYFVDIYLLMDSLYSMKDLSIQNLGKTLATQMKRLTSLN		148
ITGB4	140	YTIGFGKFDKVSVPQDMRP-EKLEKPNPNSD---PPFSFKNVLSE		184
3ije_B	149	LRIGFGAFVDPKVSFMYIISPEALENFCYDMKTTCLPMFGYKHVLTLD		198
ITGB4	185	DVDFEFRNKQGERISGNLDAPEGGFDAIQAVCTRDIGWRPDSHLLVF		234
3ije_B	199	QVTRFNEEVKQSVSRNDAPEGGFDAIQAVCTDEKIGWRNDASHLLVF		248
ITGB4	235	STESAFHYEADGANVLGIMSRNDRCHLDTGTQYQYRQDYPSPVTLV		284
3ije_B	249	TTDKATHIALDGR--LAGIVQPNQDQCHVGSNDHYSASTMDFYSLGLMT		296

ITGB4	30	RDRRCNTQAE	39
3ije_B	34	GSPRCDLKEN	43
ITGB4	30	R-----DRRCNTQAE	39
3ije_B	32	PLGSPRCDLKEN----	43

Local Alignment score drops 28



**B**



**C**

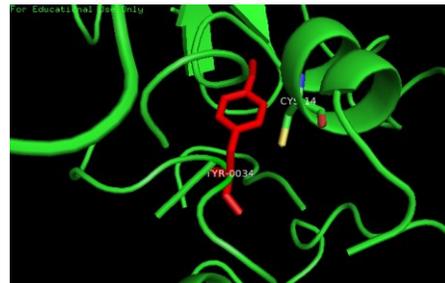


Figure 4.5 (A) Alignment comparison between original and alternative alignments of 1TGB4 and 3IJE\_B. The alternative leads to a major re-organization of local alignment and a score drop of 28, suggesting the original alignment is accurate. (B) Structural model of 1TGB4 based on the 3IJE\_B template. From the model it is clear that the substitution of CYS34 to TYR (C) abolishes a disulfide bridge between C34

and C14 and the big Tyrosine does not fit in the small space. Both of these effects are expected to contribute to destabilization of the protein.

### **Increasing the number of model-able variants by picking high confidence alignments in low sequence identity models**

The 1000 genomes project provides us a deep catalog of human variations (Durbin, Altshuler et al. 2010). This rich information gives us an unprecedented opportunity to study the distribution and impact of human genomic variation at the population level. There are more than 170000 missense SNVs. Only ~16% of these (29160) have a structure template with more than 40% sequence identity. By relaxing the sequence identity threshold to 25%, we can get many more models (total 66995), but a lot are of low quality. Using the procedure developed here, we obtained acceptable models for an additional 11030 SNVs, significantly increasing the structural coverage of SNVs from 16% to 23%.

#### 4.3: Methods

##### **Extracting pair-wise alignments from Balibase**

Balibase (Thompson, Koehl et al. 2005) multiple sequence alignments were downloaded from <http://bips.u-strasbg.fr/fr/Products/Databases/BAlIbASE/>; on 2009. The reference1 dataset was used, which contains multiple sequence alignments of equi-distant sequences. The percent identity between any two sequences within a

multiple sequence alignment is within a specific range. All the sequences within an alignment are of similar length, with no large insertions or extensions. All pair-wise alignments are extracted from a multiple alignment. For example each 6 way multiple

alignment generates  $\binom{6}{2} = 15$  pair-wise alignments.

### **Comparing Smith-Waterman alignments with extracted pair-wise alignments**

Each pair of extracted Balibase sequences was aligned using the Smith-Waterman algorithm using our program, in two ways. 1) With the BLOSUM62 substitution matrix, a gap open penalty of 10 and a gap extension penalty 0.5. 2) With the weighted sum score function (von Ohlsen, Sommer et al. 2003) score between a pair of profile positions. The formula for calculating the score is as follows:

$$W(P, Q) = \sum_{i=1}^{20} \sum_{j=1}^{20} P_i Q_j w(i, j)$$

in which  $P$  and  $Q$  are two 20 dimension vectors generated using PSI-blast against the NR database (Altschul, Madden et al. 1997).  $P_i$  is the probability of the amino-acid  $i$  at a specific position in the first sequence and  $Q_j$  is the probability of the amino-acid  $j$  at a specific position in the second sequence,  $w(i, j)$  is the substitution score for amino-acid  $i$  and  $j$  from BLOSUM62.

Then the Smith-Waterman alignment was compared with the corresponding Balibase pairwise alignment. If the Smith-Waterman algorithm aligns the same amino-acid in

the second sequence to the first sequence as in the Balibase alignment, then the position is considered to be a consistent position. Otherwise, that position is considered as an inconsistent position.

### **Force-Bypass algorithm**

Before introducing the Force-Bypass algorithm, we briefly review the Smith-Waterman algorithm. More details are in (Smith and Waterman 1981). The algorithm converts the problem of finding the highest score alignment between two sequences into a problem of finding the highest score path across a grid (Figure 4.6A), with horizontal and vertical steps denoting insertions and deletions and diagonal steps denoting matching of a pair of residues. The Smith-Waterman algorithm uses a dynamic programming procedure to find the path with the highest score across the whole grid, by finding the highest score path to every point in the grid. By a small trick of iteration, this problem is reduced to finding the smallest number among three numbers.

Let the highest score to point  $(i, j)$  in the grid be  $H(i, j)$ , we only need the highest score to its 3 adjacent points  $(i-1, j-1)$ ,  $(i-1, j)$  and  $(i, j-1)$  to calculate  $H(i, j)$  because  $H(i, j)$  must be one of these three (Figure 4.6A):

$$H(i, j) = \max \left\{ \begin{array}{l} 0 \\ H(i-1, j-1) + w(a_i, b_j) \text{ Match / Mismatch} \\ H(i-1, j) + w(a_i, -) \text{ Deletion} \\ H(i, j-1) + w(-, b_j) \text{ Insertion} \end{array} \right\}$$

To initiate the iteration, we only need to set the largest score at the border points to 0:

$$\begin{aligned} H(i, 0) &= 0, 0 \leq i \leq m \\ H(0, j) &= 0, 0 \leq j \leq n \end{aligned}$$

During the iteration, the highest score path to every point in the grid is stored. After we fill out the grid or we get to the border of the grid, we trace back each step stored during the iteration, then we get the highest score path.

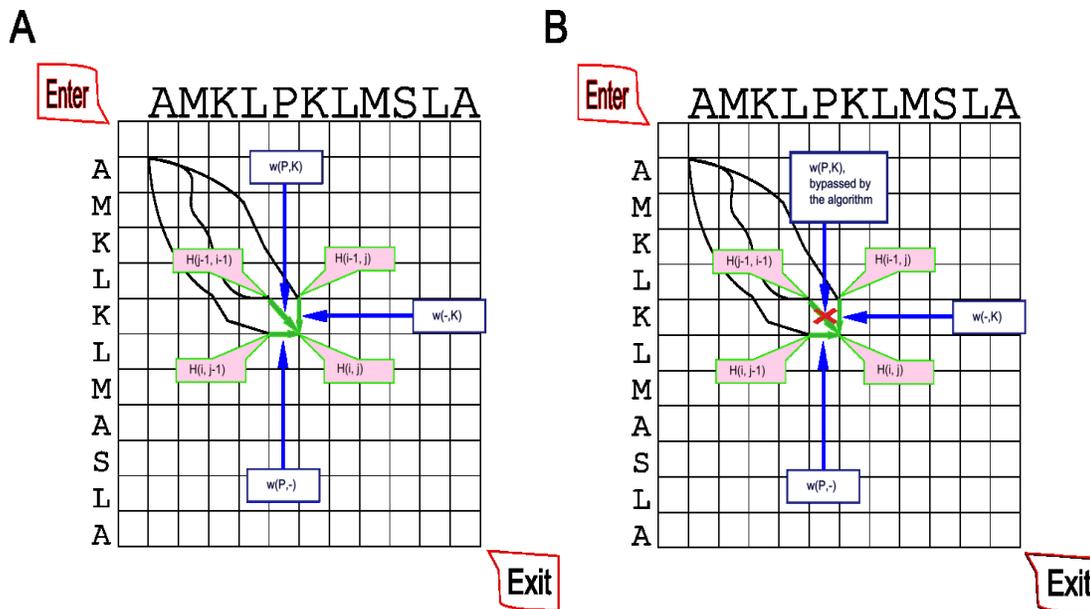


Figure 4.6 Illustration of Smith-Waterman alignment method (A) and force-bypass algorithm (B). In (A) the algorithm has three choices at the 5th position in the first sequence, it can align the Proline with Lysine in the second sequence, or introduce a gap in the first sequence or introduce a gap in the second sequence. The best choice is aligning Proline with Lysine in this case. In (B), the Proline-Lysine alignment option is excluded, since it's in the optimal alignment, the algorithm has to choose the path with highest score that doesn't align K and P in these two sequences.

We implemented the Force-Bypass algorithm within this framework, and assessed the accuracy of each position in the initial Smith-Waterman algorithm. The procedure basically uses the same iteration strategy as we described above. However, when the iteration arrives at the position we are interested in, it excludes the selection made by the Smith-Waterman algorithm, as a result, a next highest scoring path across the grid is created. Comparison of the scores of the two alignments allows us to derive a

probability that the original alignment is correct. Since the new alignment may be of a different length, and the total alignment score depends on the length of the alignment, and this may change upon application of force-bypass. To control for length differences, we only compare the scores for a fixed length 10 amino-acid fragment surrounding the position we are interested in.

### **Annotating missense mutations in the 1000 genome data**

1000 genomes project phase I data was downloaded from <http://www.1000genomes.org/data>, in May 2011. The VCF file from the website contains the position of the variations and individual genotype for more than 1000 people of various ethnic backgrounds. The REFSEQ human gene genomic position information was downloaded from the UCSC genome browser site <http://genome.ucsc.edu/> in Jan 2012. We filter out the Refseq transcripts if: (1) The coding region is not a multiple of 3; (2) The coding region doesn't start with a start codon; (3) The coding region doesn't end with a stop codon. We use those SNVs from the VCF files only when the "filter" tag equals "PASS". We end up with 177191 missense SNVs.

### **Structural modeling of missense variants**

Details of missense variant structural modeling can be found in our previous papers (Yue, Li et al. 2005; Yue, Melamud et al. 2006; Yue and Moulton 2006). Here we give a brief summary of the procedure. If an experimental structure is not available, we

choose the PDB crystal structure template with the highest sequence identity to the target protein that has a resolution better than 3Å. We copy the main chain from the template and build the side chain conformations using SCWRL (Wang, Canutescu et al. 2008; Krivov, Shapovalov et al. 2009). Given a protein structure that includes the residue substitution position, we calculate 11 protein stability features. There are four classes of electrostatic interaction: reduction of charge-charge, charge-polar or polar-polar energy, or introduction of electrostatic repulsion; three solvation effects: burying of charge or polar groups, and reduction in non-polar area buried on folding; and two terms representing steric strain: backbone strain and over-packing. The other two features considered are cavity formation (affecting van der Waals energy), and loss of a disulfide bridge. Surface accessibility of the mutated residue relative to the unfolded state is also included, as well as three parameters related to the  $C\alpha$  temperature factor of the altered residues, so that there are 15 parameters in total. We use SVMlight software <http://svmlight.joachims.org/> to determine the partitioning surface between the disease and non-disease SNVs in the 15-dimensional feature space. Continuous variables were normalized in the form of a Z score ( $Z = (\text{value} - \text{mean}) / \text{standard-deviation}$ ). A radial basis kernel with a  $g$  value of 0.2 was used with a cost factor equal to the ratio between number of negative cases and the number of positive cases.

### **Measuring the performance of variant impact modeling for different sets of models**

We used a training data set of deleterious mutations and neutral mutations, details of

which are in our previous papers (Yue, Li et al. 2005; Yue and Moulton 2006). The deleterious variants are a set of single amino acid substitutions known to cause monogenic disease. Genes associated with monogenic disease were identified by checking the Human Gene Mutation Database<sup>10</sup> (HGMD) (as of 9 February 2002) (Krawczak, Ball et al. 2000). HGMD contains a collection of mutations related to genetic diseases. Most of the mutations cause monogenic disease, although a few may be associated with disease as a result of linkage disequilibrium rather than directly causative, or contribute to a complex trait disease. Later versions of HGMD include more of the latter class, and so the earlier version was preferred and used here. For the neutral variants set, we used missense base differences between human proteins and orthologs in mouse. The justification here is that almost all variants that are fixed between species are essentially neutral and non-deleterious.

We use 30 rounds of bootstrapping to benchmark the performance of the SVM predictor trained on the 15 structural features described previously. For each round we sampled the training data with replacement, as a result 63% of the data points are sampled and treated as training data and the remaining 37% are treated as validation data. False positive and false negative rates are calculated from the validation data. We take the average of the 30 rounds as the final false positive rate and false negative rates.

We perform the benchmarking procedures on the SVMs trained on five sets of data using different sets of protein models. 1) Experimental structures and protein models

based on PDB templates of higher than 40% sequence identity to the protein of interest. 2) Protein models based on PDB templates of sequence identity between 25% and 40%, and the score drop for each mutation position calculated by our force bypass algorithm with the BLOSUM62 matrix is larger or equal to 10. 3) Protein models based on PDB templates of sequence identity between 25% and 40%, and the score drop for each mutation position calculated by our force bypass algorithm with the BLOSUM62 matrix is smaller than 10. 4) Protein models based on PDB templates of sequence identity between 25% and 40% and the score drop for each mutation position calculated by our force bypass algorithm with profile-profile scores is larger or equal to 10. 5) Protein models based on PDB templates of sequence identity between 25% and 40% and the score drop for each mutation position calculated by our force bypass algorithm with profile-profile scores is smaller than 10. Using these criteria, we get 13154 variant models for set 1, 5712 models for set 2, 7036 models for set 3, 5012 models for set 4 and 3512 models for set 5.

#### 4.4: Discussion

In this work, we developed a method to assess the accuracy of low sequence identity alignment at a single position and applied the method to SNV impact modeling. There are several pre-existing methods to assess the accuracy of alignments. For example, (Holmes and Durbin 1998) address this problem by simulating the process of random mutation for pairs of initially identical sequences to generate pseudo orthologous sequences, then compare the alignments of these simulated sequence pairs with the

supposed correct alignments. iRMSD (Armougom, Moretti et al. 2006) evaluates alignment accuracy using structure information based on the rationale that physical distance among correctly aligned amino acids should be similar in the two aligned proteins, and it requires solved protein structures. These methods do not provide the accuracy of a specific position for an alignment, crucial in applications such as SNV impact prediction.

Although it is sometimes possible to spot a bad alignment around a position by visual inspection, our method provides a more rigorous and high through-put way to do that, making it suitable for high throughput applications. As data from exome and full genome sequencing accumulate, there is an increasing need for such methods. For example, a recent exome deep re-sequencing project (Tennesen, Bigam et al. 2012) found there are a large number of rare SNVs in the human exome, most of which were previously unknown. The more high quality structure models we can build, the more we can take advantage of these data. Our method can be applied to such datasets in a similar fashion as we demonstrated using the 1000 genomes data (Durbin, Altshuler et al. 2010).

The method is not only applicable to SNV impact modeling, but also useful for evolutionary analysis, especially for those tasks requiring high accuracy alignment for low identity sequences at specific positions. For example, a recent study used comparative biology approaches to investigate the evolution of protein phosphorylation sites (Pearlman, Serber et al. 2011). Tracing the origin and evolution

of these phosphorylation sites required alignment of sequences from widely diverged organisms. The accuracy of the alignments, especially for the phosphorylation sites, is crucial for their work. Our method could be applied to these data in order to increase accuracy.

The force-bypass algorithm is implemented in a java program and free for use. We hope it will be useful for both SNV impact modeling and evolutionary analysis.

## Chapter 5: Conclusions and perspectives

In this dissertation, we demonstrated the value of computational methods in addressing a wide range of problems in protein structure, evolution, and networks. Here we give a brief summary of the conclusions for each project and discuss the future directions in each area.

### 5.1: Population Genomics: studying evolution in detail

In the first part of my dissertation, we investigated several evolutionary biology problems from the perspective of population genomics, using data from the 1000 genomes project (Durbin, Altshuler et al. 2010). First we investigated the relationship between the impact of non-synonymous SNVs and their population frequency. We found deleterious SNVs generally have low population frequencies, providing evidence of purifying selection at the population genomics level. Second, we investigated possible reasons for correlation between dN and dS using different categories of SNV densities. We found that the analogous population level measures of dN and dS, non-synonymous SNV density and synonymous SNV density, do correlate with each other and both are strongly correlated with intron SNV density. Using partial correlation analysis we showed that the correlation between the density of non-synonymous and synonymous SNVs arises both because of local mutation rate variation across the genome and shared selection pressure on these classes of SNV. We propose that the correlation between dN and dS also arises from the same two causes, mutation rate variation and shared selection pressure at the two types of site.

Third, we investigated the influence of two factors (mRNA expression level and number of protein-protein interactions) on the density of non-synonymous and synonymous SNVs and found that they both negatively correlate with the density of ns-SNVs. We also found that the correlation between predicted high impact ns-SNV density and mRNA expression level is higher than the correlation between neutral ns-SNV density and mRNA expression, which gives population level support to the mis-folding error model (Drummond, Raval et al. 2006). We also checked genes that deviate markedly from this trend and found most are immune-related., suggesting that the advantages of population diversity in these genes off-sets the fitness cost associated with increased mis-folding.

As new generation sequencing becomes cheaper and faster, there will be an explosion in deep sequencing data for human genomes. Those data will make it possible to investigate human evolution problems at a new level of detail. In this project, we focused on the evolution in coding regions, but in the future, there is much work to do on non-coding regions. For example, the human regulatory network constructed as part of the ENCODE project (Gerstein, Kundaje et al. 2012) provides us with a detailed map of human gene regulation. Integrating this data with the upcoming human sequencing data in a similar manner as we did in this project will help us understand evolution of human regulatory elements, especially in recent human history. By combined analysis of the human regulatory map, human population genomes and other functional databases, such as the human protein functional network (Wu, Feng et al. 2010) and the mammalian phenotype database (Smith,

Goldsmith et al. 2004) , we will understand the evolution of human regulatory networks at a new level. Combining the analysis of evolution for coding and noncoding regions will also let us study the interplay between the evolution of human regulatory networks and human coding sequences (Carroll 2005; Gemayel, Vincet et al. 2010), which will help in understanding genome evolution from a holistic perspective.

### 5.2: Integration of GWAS with other knowledge and the future of GWAS with rare variants

In the second part of my dissertation, I studied the ability of GWAS to discover mechanism genes through combined analysis of data from the GWAS catalog and Drugbank. We discovered that for 856 drug targets that have corresponding GWA studies, only 20 are identified. We investigated several possible reasons to explain this observation, and found two that contribute: (1) Drug target genes are strongly coupled with the disease phenotype, as a result, they have fewer common deleterious SNPs. (2) The probability of a gene being reported by GWAS is related to its transcript length and GWAS reported genes are generally longer than drug targets. Although we found GWA studies are not effective at discovering drug targets directly, we found that drug targets and GWAS reported genes are closely related in the biological network. Based on this observation, we designed a machine learning method to discover drug targets from their network relationship to GWAS reported genes.

The GWAS catalog is a public resource to facilitate the exchange and publication of experimental results, and it is useful for people with interest in specific diseases. But this centralized data resource is not been explored to its full potential when only used as a collection of independent results for different diseases. In our study, we combine the GWAS catalog with Drugbank and a Human protein functional interaction network. This allows us to understand the common properties of all the GWAS results as a whole, both their strengths and weaknesses. There are many more interesting large scale datasets for us to explore in the future. To name a few: mammalian phenotype ontology (Smith, Goldsmith et al. 2004) and Large scale literature mining data generated from many different methods (Marcotte, Xenarios et al. 2001; Becker, Hosack et al. 2003; Korb, Doerks et al. 2005; Jensen, Saric et al. 2006). These possibilities also raise challenges in natural text processing, such as we have already faced when capturing drug information for specific diseases, as well as clustering GWA studies of the same disease that have different names. If we want to integrate this information with the phenotype ontology and literature knowledge or other knowledge resources, the challenge is even bigger. To invent new algorithms or methodology for this kind of task is probably not a role computational biologists should play, but it requires us to keep up with new methods invented by our colleagues in the machine learning community (Cohen and Hunter 2004; Verspoor, Cohen et al. 2012).

GWA studies have limitations. As shown by previous studies (Clarke and Cooper ;

Stringer, Wray et al. 2011), SNPs reported in GWAS are usually common, mild effect SNPs. The introduction of next generation sequencing will gradually change this situation. The cancer genome atlas (Hampton 2006) project has established that a large diversity of missense somatic variants are involved in cancer. A large scale exome sequencing project for early onset myocardial infarction has found many rare variants in some candidate genes for cardiological diseases (Stitzel 2012), suggesting these will explain a substantial fraction of the heritability not accounted for by GWAS. As whole exome and whole genome sequencing for GWAS becomes affordable, we will be confronted with more and more rare variants. To study the relationship between rare variants with disease requires a completely different methodology from traditional GWA studies, and new statistical procedures are being developed (Price, Kryukov et al. 2010) that will help us understand the genetic basis of complex disease at a new level.

### 5.3 Protein modeling and the future of personal medicine

In the third part of my dissertation, I first studied the relationship between accuracy of pairwise protein sequence alignment and the alignment score drop using multiple alignments from Balibase as a benchmark. We used this measure to pick high accuracy alignments suitable for building high quality structural models of human missense SNV impact. We found that by adopting this strategy, we can increase the coverage of the missense SNVs from 16% to 23% without losing too much accuracy. Exome and full genome sequencing related to rare monogenic disease (Bamshad, Ng

et al. 2011), common complex disease (Stitzziel 2012), and cancer (Wei, Walia et al. 2011) (Yan, Xu et al. 2011) are all beginning to generate data containing many previously unobserved missense variants, and determining which of these contribute to the disease phenotype is an increasingly central problem. As a consequence, there are urgent challenges in this area for both experimental structural biologists and computational structural biologists. On the one hand, structural genomics people are working hard to solve more structures for every protein family (Dessailly, Nair et al. 2009). On the other hand, computational biologists are developing new methods to build better models (Bourne 2009). If experimental structures and computational models can cover all human proteins, then we can explain many of the rare variants data in a biologically meaningful way, which will not only help us understand the mechanism of complex diseases but also help us develop therapeutics in an individual specific way, which is exactly the future of personal medicine.

## Bibliography

- Abecasis, G. R., A. Auton, et al. (2012). "An integrated map of genetic variation from 1,092 human genomes." Nature **491**(7422): 56-65.
- Adzhubei, I. A., S. Schmidt, et al. (2010). "A method and server for predicting damaging missense mutations." Nat Methods **7**(4): 248-249.
- Akey, J. M., M. A. Eberle, et al. (2004). "Population history and natural selection shape patterns of genetic variation in 132 genes." PLoS biology **2**(10): e286.
- Akey, J. M., G. Zhang, et al. (2002). "Interrogating a high-density SNP map for signatures of natural selection." Genome research **12**(12): 1805-1814.
- Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.
- Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic acids research **25**(17): 3389-3402.
- Anderson, C. N., U. Ramakrishnan, et al. (2005). "Serial SimCoal: a population genetics model for data from multiple populations and points in time." Bioinformatics **21**(8): 1733-1734.
- Armougom, F., S. Moretti, et al. (2006). "The iRMSD: a local measure of sequence alignment accuracy using structural information." Bioinformatics **22**(14): e35-39.
- Baba, K., R. Shibata, et al. (2004). "Partial correlation and conditional correlation as measures of conditional independence." Australian & New Zealand Journal of Statistics **46**(4): 657-664.
- Bachtrog, D. (2004). "Evidence that positive selection drives Y-chromosome degeneration in *Drosophila miranda*." Nature genetics **36**(5): 518-522.
- Bamshad, M. J., S. B. Ng, et al. (2011). "Exome sequencing as a tool for Mendelian disease gene discovery." Nature Reviews Genetics **12**(11): 745-755.
- Becker, K. G., D. A. Hosack, et al. (2003). "PubMatrix: a tool for multiplex literature mining." BMC Bioinformatics **4**(1): 61.

- Bentley, D. R., S. Balasubramanian, et al. (2008). "Accurate whole human genome sequencing using reversible terminator chemistry." Nature **456**(7218): 53-59.
- Blount, Z. D., J. E. Barrick, et al. (2012). "Genomic analysis of a key innovation in an experimental Escherichia coli population." Nature **489**(7417): 513-518.
- Blount, Z. D., C. Z. Borland, et al. (2008). "Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli." Proc Natl Acad Sci U S A **105**(23): 7899-7906.
- Bordoli, L., F. Kiefer, et al. (2008). "Protein structure homology modeling using SWISS-MODEL workspace." Nature protocols **4**(1): 1-13.
- Bourne, P. E. (2009). Structural bioinformatics, Wiley. com.
- Bower, M. J., F. E. Cohen, et al. (1997). "Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool." Journal of molecular biology **267**(5): 1268-1282.
- Bowness, P., K. Davies, et al. (1994). "Hereditary C1q deficiency and systemic lupus erythematosus." QJM **87**(8): 455-464.
- Bradley, P., K. M. Misura, et al. (2005). "Toward high-resolution de novo structure prediction for small proteins." Science **309**(5742): 1868-1871.
- Breitkreutz, B. J., C. Stark, et al. (2008). "The BioGRID Interaction Database: 2008 update." Nucleic Acids Res **36**(Database issue): D637-640.
- Bromberg, Y. and B. Rost (2007). "SNAP: predict effect of non-synonymous polymorphisms on function." Nucleic Acids Res **35**(11): 3823-3835.
- Bucciantini, M., E. Giannoni, et al. (2002). "Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases." Nature **416**(6880): 507-511.
- Buri, P. (1956). "Gene Frequency in Small Populations of Mutant Drosophila." Evolution **10**(No.4): 367-402.
- Bustamante, C. D., A. Fledel-Alon, et al. (2005). "Natural selection on protein-coding genes in the human genome." Nature **437**(7062): 1153-1157.
- Cai, J. J., J. M. Macpherson, et al. (2009). "Pervasive hitchhiking at coding and regulatory sites in humans." PLoS genetics **5**(1): e1000336.
- Cardozo, T., M. Totrov, et al. (1995). "Homology modeling by the ICM method."

Proteins: Structure, Function, and Bioinformatics **23**(3): 403-414.

- Carley, D. W. (2005). "Drug repurposing: identify, develop and commercialize new uses for existing or abandoned drugs. Part II." IDrugs: the investigational drugs journal **8**(4): 310.
- Carroll, S. B. (2005). "Evolution at two levels: on genes and form." PLoS biology **3**(7): e245.
- Chatr-aryamontri, A., A. Ceol, et al. (2007). "MINT: the Molecular INTERaction database." Nucleic Acids Res **35**(Database issue): D572-574.
- Chen, K. and N. Rajewsky (2006). "Natural selection on human microRNA binding sites inferred from SNP data." Nature genetics **38**(12): 1452-1456.
- Chikenji, G., Y. Fujitsuka, et al. (2003). "A reversible fragment assembly method for de novo protein structure prediction." The Journal of chemical physics **119**: 6895.
- Ciriello, G., E. Cerami, et al. (2012). "Mutual exclusivity analysis identifies oncogenic network modules." Genome Res **22**(2): 398-406.
- Clarke, A. J. and D. N. Cooper "GWAS: heritability missing in action?" Eur J Hum Genet **18**(8): 859-861.
- Cohen, K. B. and L. Hunter (2004). Natural language processing and systems biology. Artificial intelligence methods and tools for systems biology, Springer: 147-173.
- Corvin, A., N. Craddock, et al. (2010). "Genome-wide association studies: a primer." Psychological medicine **40**(7): 1063.
- Crow, J. F. and M. Kimura (1970). "An introduction to population genetics theory." An introduction to population genetics theory.
- Denver, D. R., P. C. Dolan, et al. (2009). "A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes." Proceedings of the National Academy of Sciences **106**(38): 16310-16314.
- Dessailly, B. H., R. Nair, et al. (2009). "PSI-2: structural genomics to cover protein domain family space." Structure **17**(6): 869-881.
- Diniz, J., R. Almeida, et al. (2012). "Kawasaki disease and juvenile systemic lupus erythematosus." Lupus **21**(1): 89-92.
- DiStefano, J. K. and D. M. Taverna (2011). Technological issues and experimental

- design of gene association studies. Disease Gene Identification, Springer: 3-16.
- Do, R., S. Kathiresan, et al. (2012). "Exome sequencing and complex disease: practical aspects of rare variant association studies." Human molecular genetics **21**(R1): R1-R9.
- Domingos, P. (1999). MetaCost: a general method for making classifiers cost-sensitive. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.
- Drummond, D. A., A. Raval, et al. (2006). "A single determinant dominates the rate of yeast protein evolution." Mol Biol Evol **23**(2): 327-337.
- Drummond, D. A. and C. O. Wilke (2008). "Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution." Cell **134**(2): 341-352.
- Durbin, R., D. Altshuler, et al. (2010). 1000 Genomes Project: A Deep Catalog of Human Genetic Variation, January.
- Ecker, J. R., W. A. Bickmore, et al. (2012). "Genomics: ENCODE explained." Nature **489**(7414): 52-55.
- Eden, E., D. Lipson, et al. (2007). "Discovering motifs in ranked lists of DNA sequences." PLoS Comput Biol **3**(3): e39.
- Eden, E., R. Navon, et al. (2009). "GORilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists." BMC Bioinformatics **10**: 48.
- Edman, P. (1949). "A method for the determination of amino acid sequence in peptides." Arch Biochem **22**(3): 475.
- Ehret, G. B. (2010). "Genome-wide association studies: contribution of genomics to understanding blood pressure and essential hypertension." Curr Hypertens Rep **12**(1): 17-25.
- Eid, J., A. Fehr, et al. (2009). "Real-time DNA sequencing from single polymerase molecules." Science **323**(5910): 133-138.
- Eswar, N., B. John, et al. (2003). "Tools for comparative protein structure modeling and analysis." Nucleic acids research **31**(13): 3375-3380.
- Ewens, W. J. (1990). Population genetics theory-the past and the future. Mathematical and statistical developments of evolutionary theory, Springer: 177-227.

- Eyre-Walker, A. (2006). "The genomic rate of adaptive evolution." Trends Ecol Evol **21**(10): 569-575.
- Fay, J. C. and C.-I. Wu (2000). "Hitchhiking under positive Darwinian selection." Genetics **155**(3): 1405-1413.
- Fedorova, L. and A. Fedorov (2005). "Puzzles of the human genome: Why do we need our introns?" Current Genomics **6**(8): 589-596.
- Fire, M., L. Tenenboim, et al. (2011). Link prediction in social networks using computationally efficient topological features. Privacy, security, risk and trust (passat), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom), IEEE.
- Floudas, C., H. Fung, et al. (2006). "Advances in protein structure prediction and de novo protein design: A review." Chemical Engineering Science **61**(3): 966-988.
- Floyd, R. W. (1962). "Algorithm 97: shortest path." Communications of the ACM **5**(6): 345.
- Franke, A., D. P. McGovern, et al. (2010). "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci." Nature genetics **42**(12): 1118-1125.
- Fraser, H. B., A. E. Hirsh, et al. (2002). "Evolutionary rate in the protein interaction network." Science **296**(5568): 750-752.
- Fraser, H. B., D. P. Wall, et al. (2003). "A simple dependence between protein evolution rate and the number of protein-protein interactions." BMC Evolutionary Biology **3**(1): 11.
- Fu, W., T. D. O'Connor, et al. (2012). "Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants." Nature.
- Fu, Y.-X. (1995). "Statistical properties of segregating sites." Theoretical population biology **48**(2): 172-197.
- Fujita, P. A., B. Rhead, et al. (2011). "The UCSC genome browser database: update 2011." Nucleic acids research **39**(suppl 1): D876-D882.
- Gattei, V., M. Degan, et al. (1998). "Differential expression of the RET gene in human acute myeloid leukemia." Annals of hematology **77**(5): 207-210.
- Gazave, E., T. Marqués-Bonet, et al. (2007). "Patterns and rates of intron divergence

- between humans and chimpanzees." Genome biology **8**(2): R21.
- Gemayel, R., M. D. Vences, et al. (2010). "Variable tandem repeats accelerate evolution of coding and regulatory sequences." Annual review of genetics **44**: 445-477.
- Gerstein, M. B., A. Kundaje, et al. (2012). "Architecture of the human regulatory network derived from ENCODE data." Nature **489**(7414): 91-100.
- Gorlatova, N., K. Chao, et al. (2011). "Protein characterization of a candidate mechanism SNP for Crohn's disease: the macrophage stimulating protein R689C substitution." PLoS One **6**(11): e27269.
- Hampton, T. (2006). "Cancer genome atlas." JAMA: The Journal of the American Medical Association **296**(16): 1958-1958.
- Harrow, J., A. Frankish, et al. (2012). "GENCODE: the reference human genome annotation for The ENCODE Project." Genome Res **22**(9): 1760-1774.
- Hartl, D. L. and A. G. Clark (1997). Principles of population genetics, Sinauer associates Sunderland.
- Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." Proc Natl Acad Sci U S A **89**(22): 10915-10919.
- Hindorf, L. A., P. Sethupathy, et al. (2009). "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." Proc Natl Acad Sci U S A **106**(23): 9362-9367.
- Hodgkinson, A. and A. Eyre-Walker (2011). "Variation in the mutation rate across mammalian genomes." Nature Reviews Genetics **12**(11): 756-766.
- Holmes, I. and R. Durbin (1998). "Dynamic programming alignment accuracy." J Comput Biol **5**(3): 493-504.
- Hopf, T. A., L. J. Colwell, et al. (2012). "Three-dimensional structures of membrane proteins from genomic sequencing." Cell **149**(7): 1607-1621.
- Hudson, R. R. and N. L. Kaplan (1995). "Deleterious background selection with recombination." Genetics **141**(4): 1605-1617.
- Hughes, A. L. (1994). "The evolution of functionally novel proteins after gene duplication." Proceedings of the Royal Society of London. Series B: Biological Sciences **256**(1346): 119-124.
- Hughes, A. L., B. Packer, et al. (2003). "Widespread purifying selection at

- polymorphic sites in human protein-coding loci." Proceedings of the National Academy of Sciences **100**(26): 15754-15757.
- Hung, L.-H., S.-C. Ngan, et al. (2007). De novo protein structure prediction. Computational methods for protein structure prediction and modeling, Springer: 43-63.
- Imanishi, T. and H. Nakaoka (2009). "Hyperlink Management System and ID Converter System: enabling maintenance-free hyperlinks among major biological databases." Nucleic Acids Res **37**(Web Server issue): W17-22.
- Imming, P., C. Sinning, et al. (2006). "Drugs, their targets and the nature and number of drug targets." Nat Rev Drug Discov **5**(10): 821-834.
- Jensen, L. J., J. Saric, et al. (2006). "Literature mining for the biologist: from information retrieval to biological discovery." Nature Reviews Genetics **7**(2): 119-129.
- Jia, P., L. Wang, et al. (2012). "Network-assisted investigation of combined causal signals from genome-wide association studies in schizophrenia." PLoS Comput Biol **8**(7): e1002587.
- Jia, P. and Z. Zhao (2011). "Network-assisted Causal Gene Detection in Genome-wide Association Studies: An Improved Module Search Algorithm." IEEE Int Workshop Genomic Signal Process Stat: 131-134.
- Johnson, A. D., C. Newton-Cheh, et al. (2011). "Association of hypertension drug target genes with blood pressure and hypertension in 86,588 individuals." Hypertension **57**(5): 903-910.
- Johnson, R. A. and D. W. Wichern (2002). Applied multivariate statistical analysis, Prentice hall Upper Saddle River, NJ.
- Jordan, I. K., Y. I. Wolf, et al. (2003). "No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly." BMC Evolutionary Biology **3**(1): 1.
- Jorde, L. B. (1995). "Linkage disequilibrium as a gene-mapping tool." American journal of human genetics **56**(1): 11.
- Kanehisa, M. (2002). "The KEGG database." Novartis Found Symp **247**: 91-101; discussion 101-103, 119-128, 244-152.
- Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic Acids Res **28**(1): 27-30.

- Kanehisa, M., S. Goto, et al. (2002). "The KEGG databases at GenomeNet." Nucleic Acids Res **30**(1): 42-46.
- Kanehisa, M., S. Goto, et al. (2004). "The KEGG resource for deciphering the genome." Nucleic Acids Res **32**(Database issue): D277-280.
- Karplus, K. and B. Hu (2001). "Evaluation of protein multiple alignments by SAM-T99 using the BALiBASE multiple alignment test set." Bioinformatics **17**(8): 713-720.
- Kiezun, A., K. Garimella, et al. (2012). "Exome sequencing and the genetic basis of complex traits." Nature genetics **44**(6): 623-630.
- Kim, S. H. and S. V. Yi (2006). "Correlated asymmetry of sequence and functional divergence between duplicate proteins of *Saccharomyces cerevisiae*." Mol Biol Evol **23**(5): 1068-1075.
- Kimura, M. (1977). "Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution." Nature **267**(5608): 275-276.
- King, M.-C., J. H. Marks, et al. (2003). "Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2." Science **302**(5645): 643-646.
- Knox, C., V. Law, et al. (2011). "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs." Nucleic Acids Res **39**(Database issue): D1035-1041.
- Koonin, E. V., K. S. Makarova, et al. (2001). "Horizontal gene transfer in prokaryotes: quantification and classification 1." Annual Reviews in Microbiology **55**(1): 709-742.
- Korb, L. C. and J. M. Ahearn (1997). "C1q binds directly and specifically to surface blebs of apoptotic human keratinocytes: complement deficiency and systemic lupus erythematosus revisited." The Journal of Immunology **158**(10): 4525-4528.
- Korbel, J. O., T. Doerks, et al. (2005). "Systematic association of genes to phenotypes by genome and literature mining." PLoS biology **3**(5): e134.
- Krawczak, M., E. V. Ball, et al. (2000). "Human gene mutation database—a biomedical information and research resource." Hum Mutat **15**(1): 45-51.
- Krivov, G. G., M. V. Shapovalov, et al. (2009). "Improved prediction of protein side-chain conformations with SCWRL4." Proteins **77**(4): 778-795.
- Kryazhimskiy, S. and J. B. Plotkin (2008). "The population genetics of dN/dS." PLoS

genetics **4**(12): e1000304.

- Krylov, D. M., Y. I. Wolf, et al. (2003). "Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution." Genome Res **13**(10): 2229-2235.
- Kryshtafovych, A., K. Fidelis, et al. (2011). "CASP9 results compared to those of previous CASP experiments." Proteins: Structure, Function, and Bioinformatics **79**(S10): 196-207.
- Kulseth, M. A., K. E. Berge, et al. (2010). "Analysis of LDLR mRNA in patients with familial hypercholesterolemia revealed a novel mutation in intron 14, which activates a cryptic splice site." Journal of human genetics **55**(10): 676-680.
- Kumar, A., T. A. White, et al. (2011). "Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers." Proceedings of the National Academy of Sciences **108**(41): 17087-17092.
- Lang, M., M. Stelzer, et al. "BKM-react, an integrated biochemical reaction database." BMC Biochem **12**: 42.
- Laxer, R., B. Cameron, et al. (1988). "Occurrence of Kawasaki disease and systemic lupus erythematosus in a single patient." The Journal of rheumatology **15**(3): 515.
- Lee, H. K., A. K. Hsu, et al. (2004). "Coexpression analysis of human genes across many microarray data sets." Genome Res **14**(6): 1085-1094.
- Lee, J., A. Liwo, et al. (1999). "Energy-based de novo protein folding by conformational space annealing and an off-lattice united-residue force field: Application to the 10-55 fragment of staphylococcal protein A and to apo calbindin D9K." Proceedings of the National Academy of Sciences **96**(5): 2025-2030.
- Lemos, B., B. R. Bettencourt, et al. (2005). "Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions." Mol Biol Evol **22**(5): 1345-1354.
- Li, W. H., C. I. Wu, et al. (1985). "A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes." Mol Biol Evol **2**(2): 150-174.
- Lynch, M. (2007). "The origins of genome architecture."
- Lynch, M. and J. S. Conery (2003). "The origins of genome complexity." Science

**302**(5649): 1401-1404.

- Marcotte, E. M., I. Xenarios, et al. (2001). "Mining literature for protein–protein interactions." Bioinformatics **17**(4): 359-363.
- Marth, G. T., F. Yu, et al. (2011). "The functional spectrum of low-frequency coding variation." Genome Biol **12**(9): R84.
- Matthews, L., G. Gopinath, et al. (2009). "Reactome knowledgebase of human biological pathways and processes." Nucleic Acids Res **37**(Database issue): D619-622.
- McDonald, J. H. and M. Kreitman (1991). "Adaptive protein evolution at the Adh locus in Drosophila." Nature **351**(6328): 652-654.
- McDonald, M. J., W.-C. Wang, et al. (2011). "Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences." PLoS biology **9**(6): e1000622.
- Melamud, E. and J. Moulton (2009). "Stochastic noise in splicing machinery." Nucleic Acids Res **37**(14): 4873-4886.
- Metz, E. C., R. Robles-Sikisaka, et al. (1998). "Nonsynonymous substitution in abalone sperm fertilization genes exceeds substitution in introns and mitochondrial DNA." Proc Natl Acad Sci U S A **95**(18): 10676-10681.
- Moulton, J. (2008). "Comparative modeling in structural genomics." Structure **16**(1): 14-15.
- Nachman, M. W., S. N. Boyer, et al. (1994). "Nonneutral evolution at the mitochondrial NADH dehydrogenase subunit 3 gene in mice." Proceedings of the National Academy of Sciences **91**(14): 6364-6368.
- Nelson, M. R., D. Wegmann, et al. (2012). "An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people." Science **337**(6090): 100-104.
- Ng, P. C. and S. Henikoff (2003). "SIFT: Predicting amino acid changes that affect protein function." Nucleic Acids Res **31**(13): 3812-3814.
- Nik-Zainal, S., L. B. Alexandrov, et al. (2012). "Mutational processes molding the genomes of 21 breast cancers." Cell.
- Ohashi, J. and K. Tokunaga (2001). "The power of genome-wide association studies of complex disease genes: statistical limitations of indirect approaches using SNP markers." Journal of human genetics **46**(8): 478-482.

- Ohlson, T., B. Wallner, et al. (2004). "Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods." Proteins **57**(1): 188-197.
- Ohta, T. and Y. Ina (1995). "Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences." J Mol Evol **41**(6): 717-720.
- Overington, J. P., B. Al-Lazikani, et al. (2006). "How many drug targets are there?" Nat Rev Drug Discov **5**(12): 993-996.
- Pal, C., B. Papp, et al. (2006). "An integrated view of protein evolution." Nat Rev Genet **7**(5): 337-348.
- Parenteau, J., M. Durand, et al. (2008). "Deletion of many yeast introns reveals a minority of genes that require splicing for function." Molecular biology of the cell **19**(5): 1932-1941.
- Parsch, J., S. Novozhilov, et al. (2010). "On the utility of short intron sequences as a reference for the detection of positive and negative selection in Drosophila." Mol Biol Evol **27**(6): 1226-1234.
- Pearlman, S. M., Z. Serber, et al. (2011). "A mechanism for the evolution of phosphorylation sites." Cell **147**(4): 934-946.
- Pegueroles, C., S. Laurie, et al. (2013). "Accelerated evolution after gene duplication: a time-dependent process affecting just one copy." Mol Biol Evol.
- Pleasance, E. D., R. K. Cheetham, et al. (2010). "A comprehensive catalogue of somatic mutations from a human cancer genome." Nature **463**(7278): 191-196.
- Pleasance, E. D., P. J. Stephens, et al. (2010). "A small-cell lung cancer genome with complex signatures of tobacco exposure." Nature **463**(7278): 184-190.
- Polley, S. D. and D. J. Conway (2001). "Strong diversifying selection on domains of the Plasmodium falciparum apical membrane antigen 1 gene." Genetics **158**(4): 1505-1512.
- Price, A. L., G. V. Kryukov, et al. (2010). "Pooled association tests for rare variants in exon-resequencing studies." Am J Hum Genet **86**(6): 832-838.
- Prieto, C., A. Risueno, et al. (2008). "Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles." PLoS One **3**(12): e3911.

- Pruitt, K. D., T. Tatusova, et al. (2007). "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucleic acids research **35**(suppl 1): D61-D65.
- Ramani, A. K. and E. M. Marcotte (2003). "Exploiting the co-evolution of interacting proteins to discover interaction specificity." Journal of molecular biology **327**(1): 273-284.
- Ratajewski, M., H. de Bousac, et al. (2012). "ABCC6 expression is regulated by CCAAT/enhancer-binding protein activating a primate-specific sequence located in the first intron of the gene." Journal of Investigative Dermatology.
- Reetz, M. T., L. W. Wang, et al. (2006). "Directed Evolution of Enantioselective Enzymes: Iterative Cycles of CASTing for Probing Protein-Sequence Space." Angewandte Chemie **118**(8): 1258-1263.
- Rocha, E. P. and A. Danchin (2004). "An analysis of determinants of amino acids substitution rates in bacterial proteins." Mol Biol Evol **21**(1): 108-116.
- Rockman, M. V., S. S. Skrovaneck, et al. (2010). "Selection at linked sites shapes heritable phenotypic variation in *C. elegans*." Science **330**(6002): 372-376.
- Rohl, C. A. and D. Baker (2002). "De novo determination of protein backbone structure from residual dipolar couplings using Rosetta." Journal of the American Chemical Society **124**(11): 2723-2729.
- Rossin, E. J., K. Lage, et al. (2011). "Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology." PLoS Genet **7**(1): e1001273.
- Rothberg, J. M., W. Hinz, et al. (2011). "An integrated semiconductor device enabling non-optical genome sequencing." Nature **475**(7356): 348-352.
- Rychlewski, L., L. Jaroszewski, et al. (2000). "Comparison of sequence profiles. Strategies for structural predictions using sequence information." Protein Sci **9**(2): 232-241.
- Sali, A. and T. Blundell (1994). "Comparative protein modelling by satisfaction of spatial restraints." Protein structure by distance analysis **64**: C86.
- Schneider, S., D. Roessli, et al. (2000). "Arlequin: a software for population genetics data analysis." User manual ver 2: 2496-2497.
- Sharp, P. M. and W. H. Li (1987). "The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias." Mol Biol Evol **4**(3): 222-230.

- Shi, Z. and J. Moulton (2011). "Structural and functional impact of cancer-related missense somatic mutations." Journal of molecular biology **413**(2): 495-512.
- Shriner, D., R. Shankarappa, et al. (2004). "Influence of random genetic drift on human immunodeficiency virus type 1 env evolution during chronic infection." Genetics **166**(3): 1155-1164.
- Smith, C. L., C.-A. W. Goldsmith, et al. (2004). "The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information." Genome biology **6**(1): R7.
- Smith, N. G. and A. Eyre-Walker (2002). "Adaptive protein evolution in *Drosophila*." Nature **415**(6875): 1022-1024.
- Smith, T. F. and M. S. Waterman (1981). "Identification of common molecular subsequences." J Mol Biol **147**(1): 195-197.
- Sober, S., E. Org, et al. (2009). "Targeting 160 candidate genes for blood pressure regulation with a genome-wide genotyping array." PLoS One **4**(6): e6034.
- Stahl, E. A., S. Raychaudhuri, et al. (2010). "Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci." Nature genetics **42**(6): 508-514.
- Stenson, P. D., E. V. Ball, et al. (2009). "The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalised genomics." Human genomics **4**(2): 69.
- Stenson, P. D., E. V. Ball, et al. (2003). "Human Gene Mutation Database (HGMD): 2003 update." Hum Mutat **21**(6): 577-581.
- Stephens, J. C., J. A. Schneider, et al. (2001). "Haplotype variation and linkage disequilibrium in 313 human genes." Science **293**(5529): 489-493.
- Stitzel, N. O. (2012). "RARE CODING MUTATIONS AND RISK FOR EARLY-ONSET MYOCARDIAL INFARCTION: AN EXOME SEQUENCING STUDY OF > 2,000 CASES AND CONTROLS." Journal of the American College of Cardiology **59**(13s1): E435-E435.
- Stringer, S., N. R. Wray, et al. (2011). "Underestimated effect sizes in GWAS: fundamental limitations of single SNP analysis for dichotomous phenotypes." PLoS One **6**(11): e27964.
- Su, A. I., T. Wiltshire, et al. (2004). "A gene atlas of the mouse and human protein-encoding transcriptomes." Proceedings of the National Academy of Sciences

of the United States of America **101**(16): 6062-6067.

- Subramanian, S. and S. Kumar (2004). "Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome." Genetics **168**(1): 373-381.
- Sugden, A. M., B. R. Jasny, et al. (2003). "Charting the evolutionary history of life." Science **300**(5626): 1691-1691.
- Tajima, F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." Genetics **123**(3): 585-595.
- Takahashi, N. and O. Smithies (2004). "Human genetics, animal models and computer simulations for studying hypertension." Trends Genet **20**(3): 136-145.
- Tennessen, J. A., A. W. Bigham, et al. (2012). "Evolution and functional impact of rare coding variation from deep sequencing of human exomes." Science **337**(6090): 64-69.
- Thompson, J. D., P. Koehl, et al. (2005). "BALiBASE 3.0: latest developments of the multiple sequence alignment benchmark." Proteins **61**(1): 127-136.
- Thompson, J. R., J. Attia, et al. (2011). "The meta-analysis of genome-wide association studies." Briefings in bioinformatics **12**(3): 259-269.
- Tian, D., Q. Wang, et al. (2008). "Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes." Nature **455**(7209): 105-108.
- Verspoor, K., K. B. Cohen, et al. (2012). "A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools." BMC Bioinformatics **13**(1): 207.
- von Ohlsen, N., I. Sommer, et al. (2003). "Profile-profile alignment: a powerful tool for protein structure prediction." Pac Symp Biocomput: 252-263.
- Walport, M. J., K. A. Davies, et al. (1998). "C1q and systemic lupus erythematosus." Immunobiology **199**(2): 265-285.
- Wang, K., J. Kan, et al. (2011). "Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer." Nature genetics **43**(12): 1219-1223.
- Wang, Q., A. A. Canutescu, et al. (2008). "SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling." Nat Protoc **3**(12): 1832-1847.

- Waterman, M. S. (1983). "Sequence alignments in the neighborhood of the optimum with general application to dynamic programming." Proc Natl Acad Sci U S A **80**(10): 3123-3124.
- Webster, A. J., R. J. Payne, et al. (2003). "Molecular phylogenies link rates of evolution and speciation." Science **301**(5632): 478-478.
- Wei, X., V. Walia, et al. (2011). "Exome sequencing identifies GRIN2A as frequently mutated in melanoma." Nature genetics **43**(5): 442-446.
- Witten, I. H., E. Frank, et al. (1999). "Weka: Practical machine learning tools and techniques with Java implementations."
- Wolfe, K. H. and P. M. Sharp (1993). "Mammalian gene evolution: nucleotide sequence divergence between mouse and rat." Journal of molecular evolution **37**(4): 441-456.
- Wright, S. I., I. V. Bi, et al. (2005). "The effects of artificial selection on the maize genome." Science **308**(5726): 1310-1314.
- Wright, S. I., C. B. Yau, et al. (2004). "Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*." Mol Biol Evol **21**(9): 1719-1726.
- Wu, G., X. Feng, et al. (2010). "A human functional protein interaction network and its application to cancer data analysis." Genome Biol **11**(5): R53.
- Wuchty, S. (2004). "Evolution and topology in the yeast protein interaction network." Genome research **14**(7): 1310-1314.
- Yamasaki, C., K. Murakami, et al. (2008). "The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts." Nucleic acids research **36**(Database issue): D793.
- Yampolsky, L. Y. and A. Stoltzfus (2005). "The exchangeability of amino acids in proteins." Genetics **170**(4): 1459-1472.
- Yan, X.-J., J. Xu, et al. (2011). "Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia." Nature genetics **43**(4): 309-315.
- Yeh, S. H., H. Y. Yeh, et al. (2012). "A network flow approach to predict drug targets from microarray data, disease genes and interactome network - case study on prostate cancer." J Clin Bioinforma **2**(1): 1.

- Yona, G. and M. Levitt (2002). "Within the twilight zone: a sensitive profile-profile comparison tool based on information theory." J Mol Biol **315**(5): 1257-1275.
- Yue, P., Z. Li, et al. (2005). "Loss of protein structure stability as a major causative factor in monogenic disease." J Mol Biol **353**(2): 459-473.
- Yue, P., E. Melamud, et al. (2006). "SNPs3D: candidate gene and SNP selection for association studies." BMC Bioinformatics **7**: 166.
- Yue, P. and J. Moulton (2006). "Identification and analysis of deleterious human SNPs." J Mol Biol **356**(5): 1263-1274.
- Zeggini, E. and J. P. Ioannidis (2009). "Meta-analysis in genome-wide association studies." Pharmacogenomics **10**(2): 191-201.
- Zeng, K., Y.-X. Fu, et al. (2006). "Statistical tests for detecting positive selection by utilizing high-frequency variants." Genetics **174**(3): 1431-1439.
- Zhao, Z., Y.-X. Fu, et al. (2003). "Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution." Gene **312**: 207-213.