INTERACTOME-SCALE INTERROGATIONS OF HUMAN GENOMIC VARIATION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Robert Fragoza

May 2018

© 2018 Robert Fragoza

INTERACTOME-SCALE INTERROGATIONS OF HUMAN GENOMIC VARIATION

Robert Fragoza, Ph. D.

Cornell University 2018

Coding variants segregating in human populations are expected to be largely benign, with deleterious variation occurring principally at rare allele frequencies and limited to conserved genomic sites. The extent to which this deleterious variation burdens human genomes and the mechanisms by which these mutations exert their function, though, remains largely unexplored. To help address this gap, I have contributed towards the development of interactome-scale tools for interrogating missense variation in human disease as well as experimentally measured the impact of thousands of human missense variants on protein interactions and stability. The accumulation of these efforts have helped in characterizing molecular mechanisms of disease-associated mutations and have enabled new insights towards the extent to which functional variation segregates across different human populations.

To begin, the development of a massively parallel, site-directed mutagenesis platform for cloning DNA variants, named Clone-seq, is discussed. A study of the impact of 204 disease-associated mutations on protein interactions and stability is then detailed to demonstrate the utility of Clone-seq in genomic studies. Next, an extensive study of >2,000 missense mutations is presented in which widespread protein interaction perturbations by both rare and common human population variants is

unveiled. Disruptive variants were found to be enriched within conserved sites in the genome and occurred at increasingly higher rates as allele frequency decreased. Evidence suggesting that disruptive variants persist primarily in less essential regions of the genome is then presented followed by a demonstration of how shared interaction perturbation profiles between population variants and disease-associated mutations can be applied to identify candidate disease-associated mutations from sequencing data. Lastly, the development of an integrated computational and experimental platform for prioritizing *de novo* missense mutations in developmental disorders is discussed.

While protein interaction perturbations represent only one of a multitude of ways in which DNA variants can alter cellular function, nonetheless, the genetic, protein interaction, and population-level insights presented here should represent an important step forward towards an improved understanding of the evolutionary forces that shape the human genome and protein function.

BIOGRAPHICAL SKETCH

Robert Fragoza was born June 23, 1989 in Los Angeles, California. He was the fifth of six children to Salvador and Belen Fragoza. Before grade school, Robert's family moved to Amado, Arizona, a small town of less than 1,000 people near the US-Mexico border. Robert was educated entirely within the Sahuarita Unified School District where he received numerous opportunities to explore his ever-changing interests, including participating in his elementary school choir, writing for his school's newspaper and participating in the Science Olympiad in middle school, as well as joining the boys' tennis team and being recognized as a Leader in Character for Sahuarita High School, an honor that his mother jokingly dismissed because his niece had recently received the same district-wide honor only a few months prior. Motivated by his upstanding high school chemistry teacher, Karin Rojahn, to pursue a rigorous college degree, Robert returned to Los Angeles to attend the University of Southern California where he graduated with academic honors with a B.S. in Chemical Engineering. Encouraged again by a nurturing mentor, Dr. Steven Finkle, Robert then traveled to Ithaca, New York to pursue a graduate degree in molecular biology from Cornell University. Robert now eagerly contemplates a return once again to the west coast to be closer to his family and to pursue his next career challenge.

To my extraordinary brothers and sisters, no accomplishment is made alone

ACKNOWLEDGMENTS

First and foremost, I would like to thank my scientific mentor and thesis advisor, Dr. Haiyuan Yu. Throughout my graduate studies, Haiyuan has always pushed me to pursue the highest caliber of scientific research without ever sacrificing even an ounce of quality or integrity. I know Haiyuan would never push a student who he felt could not achieve his rigorous scientific standards and for that I am thankful. His lab has not only given me an opportunity to grow as a scientist and academic, but also an opportunity to grow as an individual, the importance of which cannot be measured. I also sincerely appreciate the adventurous spirit his lab has always embraced. While my thesis focuses on large-scale cloning, protein interactions, and human coding variation, the lab has dipped its feet into the diverse realms of cross-linking mass spectrometry, gene transcription, autism and infertility risk, cancer genomics, and even a summer spent building a microfluidic device for on-chip detection of protein-protein interactions. While my success in the last category listed was quite limited, the boldness to pursue a device completely outside of our expertise or familiarity is an attitude I will keep with me in my future pursuits.

I would also like to thank my wonderfully diverse and talented fellow lab members who have helped me grow as a scientist and become cherished friends. I thank Tommy Vo whose professionalism and determination holds my sincere admiration. I thank Jishnu Das whose prodigious scientific acumen is seconded only by his strength of character. I thank Nicolas Cordero whose compassion for others has served as a permanent template for how I try to govern myself around all others. I also

wish to thank Jin Liang and Xiaomu Wei whose expert scientific advice has been invaluable towards my growth as a scientist. I also owe thanks to Michael Meyer and Juan Beltran whose answers to my mundane programming questions have made me an infinitely better coder as well as Siwei Chen, Charles Liang, and Shayne Wierbowski whose dedicated efforts helped propel my projects to success even when my own efforts would start to wane. I owe debts to all of you.

Above all else, though, I would like to thank each of my siblings. When my father couldn't make it to my college freshman orientation, my brother Eloy accompanied me instead. Trapped in Chicago for a few days due to my own carelessness and poor planning, my sister Susy got me home, no questions asked. Broke after a long semester but with a childish yearning to see USC play in the Rose Bowl, my brother Sal bought me tickets and joined me at the game (only to watch his precious Illinois get trounced 49-17). Stuck in Mexico and apathetic towards meeting my grandparents, my sister Lorena reminded me why I should instead be appreciative. To my sister Vanessa, how you put up with me during my irritable years as a moody teenager (and even more embarrassingly as a moody adult), I will never know.

Lastly, to my parents Belen and Salvador Fragoza. No text, no matter how thoughtful or well-composed, could ever express my gratitude to you both. Thank you for everything.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

In regards to the overall organization of this dissertation, Chapter 1 begins with a short introduction to basic population genetic concepts of random genetic drift and selection. These concepts are introduced to familiarize the reader with the basic rationale used by population geneticists to explain how explosive population growth in the human population can result in an influx of genetic variation and how it is that deleterious variation can still persist in human populations by random genetic drift despite purifying selection acting to purge such variation. Following this primer on genetic drift and selection, human interactome networks are then introduced to provide the reader with the necessary background to understand (1) how disease phenotypes can be interpreted as perturbations within protein interaction networks and (2) what experimental and computational tools are available for studying and constructing interactome networks. Lastly, two sections regarding recent, high profile literature on large-scale sequencing efforts are presented to familiarize the reader with how such sequencing studies are performed, what these studies reveal about functional variation in human genomes, and what the limitations of these studies are.

In Chapter 2, my research efforts towards the construction of a large-scale site-directed mutagenesis platform, Clone-seq, is discussed. A study of the impact of 204 disease-associated mutations on protein stability and interactions using large-scale GFP and yeast two-hybrid assays, respectively, is also discussed. Chapter 3 follows with a large-scale effort to measure the impact of >2,000 missense variants on protein-protein interactions and its implications towards human genomes. Particular focus is

paid on where disruptive variation propagates within human genomes and how shared interaction perturbation profiles between population variants and disease-associated mutations can be used to identify candidate disease-associated mutations. Chapter 4 then discusses the development of an integrated experimental and computational platform for identifying and prioritizing functional variation in developmental disorder sequencing studies through the use of an interactome perturbation framework. Lastly, in Chapter 5, an overall summary of the entire thesis is presented, ending with an overall perspective on how studies such as those presented here will shape our understanding of the impact genetic variation on human health and protein function.

CHAPTER 1

AN INTRODUCTION TO POPULATION GENETICS AND SEMINAL FINDINGS IN LARGE-SCALE SEQUENCING STUDIES

1.1 Genetic drift and fixation

The frequency of an allele will change due to stochastic variation in the reproductive success of individuals in a species. If a spontaneous mutation has no measurable impact on an organism's fitness, then the frequency of a particular allele across future generations should be entirely dictated by random genetic drift. We apply the Wright-Fisher model to develop a working understanding of the fundamental mechanics by which an allele entirely subject to random genetic drift can still reach fixation or become lost.

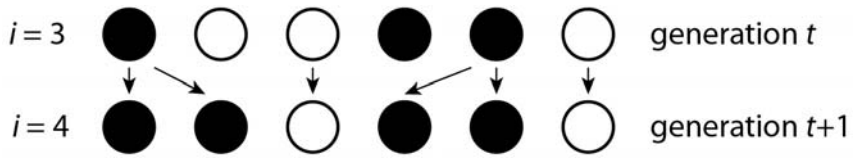


Figure 1.1-1 *Wright-Fisher model for population $N = 6$ demonstrating random fluctuation of an allele i (black-filled circle) across a single generation.*

In a Wright-Fisher model, we assume a haploid population of fixed size in which a particular allele, i, is initially present in half the population ($i = 3$ in **Figure 1.1-1**). Random mating is also assumed such that any individual in a Wright-Fisher population is equally likely to be the parent of an individual in the proceeding generation. The size of the population, $N = 6$, stays constant across generations, and the expected allele frequency i across each successive generation is shown below:

$$E[i_t|i_{t-1}] = N(i_{t-1}/N) \quad (1.1\text{-}1)$$

Hence, we expect the frequency of allele i to stay constant across successive generations; however, random mating introduces variation to this expectation. Consequently, over a large enough number of successive generations, allele i will always either become fixed in the population or become lost (**Figure 1.1-2**).

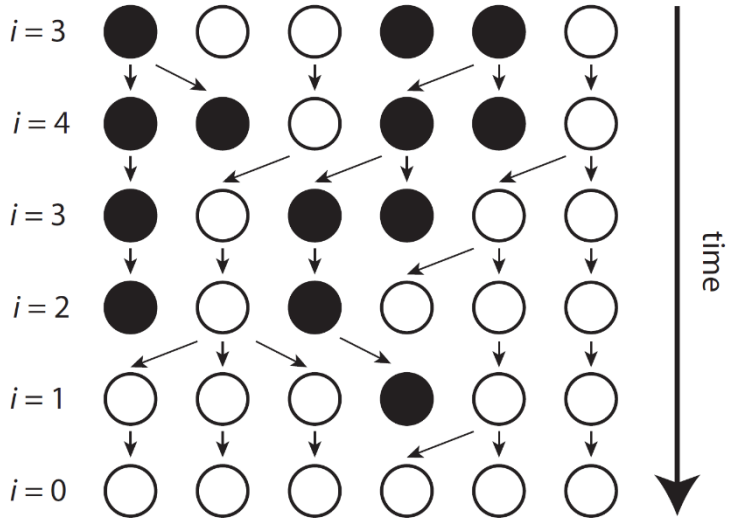


Figure 1.1-2 *In a Wright-Fisher population, random genetic drift can result in the loss of an allele ($i = 0$) over time. Though not shown here, fixation of an allele ($i = 6$) is equally likely.*

The Wright-Fisher model is therefore an excellent demonstration of the dynamic effects that random genetic drift has on the frequency of alleles across generations. Note, though, that not all mutations have a neutral impact on organism fitness. As such, we must also examine the extensive influence of selection on allele frequency.

1.2 Selection and adaptation

Adaptation to ever-changing environmental conditions requires a large and genetically diverse population in which a mutation-derived beneficial trait can either sweep through a population or, conversely, a mutation-derived deleterious trait can be readily

purged from a population. In diploid organisms, such heritable adaptations are propagated or diminished through sexual selection. As a simple demonstration of how selection can be modeled in context with random genetic drift, we first define the concept of *absolute fitness* as follows:

$$E[N_A(t+1)] = \omega_A \cdot N_A(t) \tag{1.2-1}$$

where N_A is the number of individuals that carry allele A in generation t and ω_A is a coefficient quantifying the fitness advantage/disadvantage that allele A confers upon individuals carrying this allele. Note that for a neutral allele, Equation 1.2-1 is essentially equivalent to Equation 1.1-1 in which random genetic drift is modeled.

To quantify the relative fitness advantage/disadvantage of allele A in comparison to a competing allele at the same locus, a, we define *relative fitness* as follows:

$$\frac{\omega_A}{\omega_a} = 1 + s_{Aa} \tag{1.2-2}$$

where s_{Aa} is the selection coefficient for allele A over a. In this regard, the selection coefficient s_{Aa} quantifies the relative advantage or disadvantage, if any, of carrying allele A instead of a. More specifically, we discuss the value of the selection coefficient s_{Aa} with respect to three regimes:

$$\begin{aligned} &s_{Aa} > 0, \;\; allele\ A\ is\ advantageous\ over\ a \\ &s_{Aa} = 0, \;\; allele\ A\ is\ neutral\ with\ a \\ &s_{Aa} < 0, \;\; allele\ A\ is\ disadvantageous\ over\ a \end{aligned} \tag{1.2-3}$$

As an example, we note that a mutation with $s_{Aa} = 0.2$ implies that individuals carrying allele A will grow in frequency by an average of 20% per generation. We can also model this projected growth by applying Equations 1.2-1 and 1.2-2 to the

following equation for calculating allele frequency as a function of time, $x(t)$:

$$x(t) = \frac{N_A(t)}{N_A(t) + N_a(t)} \tag{1.2-4}$$

Applying Equation 1.2-1 and 1.2-2, detailed in Appendix A.2, yields the following:

$$x(t) = \frac{x_0}{x_0 + (1 - x_0)e^{-st}} \tag{1.2-5}$$

where x_0 is the initial allele frequency at t_0 and the selection coefficient is represented as s for convenience. We note that Equation 1.2-5 is logistic and hence the frequency of alleles subject only to selection, and not random genetic drift, will demonstrate logistic growth as shown in Figure 1.2-1 under varying selection coefficients.

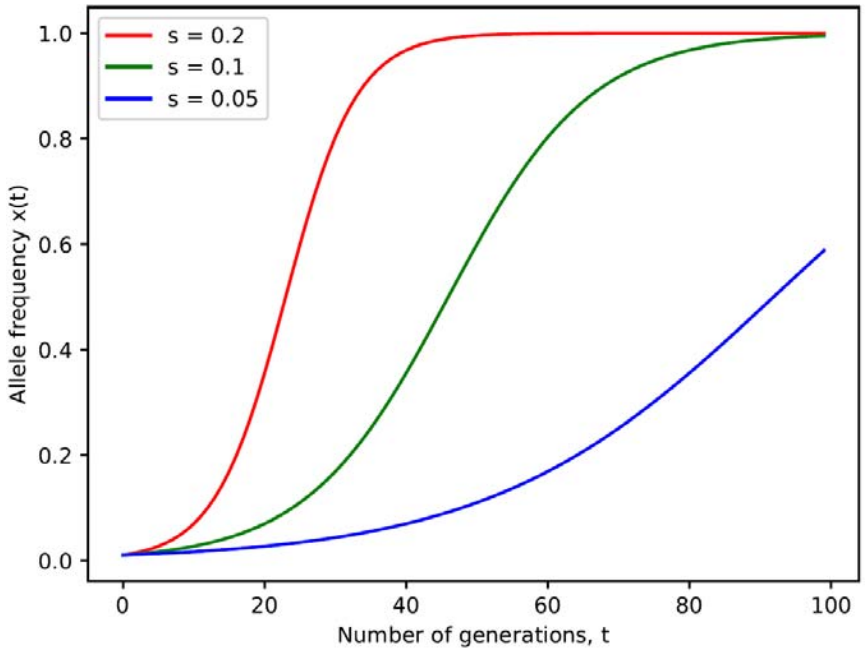


Figure 1.2-1 Variants exclusively experiencing positive selection experience logistic growth. *For all three simulations,* $x_0 = 1/100$ *indicating a de novo mutation under three decreasingly advantageous selection coefficients:* $s = 0.2, 0.1, and\ 0.05$.

To demonstrate how random genetic drift influences selection, consider the following scenario in which a newly arisen allele, $i = 1$ shown in black in **Figure 1.2-**

2, is present in the preceding generation, t, but fails to be transmitted to the proceeding generation, $t + 1$.

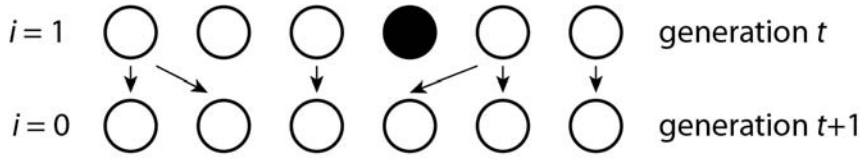


Figure 1.2-2 A single black allele in generation t fails to be transmitted to generation $t + 1$.

We can calculate the probability that no individual in generation $t + 1$ is parented by the individual carrying the black allele under a Wright-Fisher model assuming a haploid, constant population size. Under a scenario in which only random genetic drift acts, we note that the probability that allele i never reproduces is essentially a series of Bernoulli trials in which the probability of the black allele being selected is $1/6$. Therefore, the probability that no individual in generation $t + 1$ is parented by the individual carrying the black allele is:

$$\begin{aligned} \Pr(Allele\ i\ is\ lost\ in\ next\ generation) &= \left(1 - \frac{N_i}{N}\right)^N \\ &= \left(1 - \frac{1}{6}\right)^6 = 0.335 \end{aligned} \tag{1.2-6}$$

Under this scenario, there is an approximately one in three chance that the newly arisen allele does not propagate to the next generation. Suppose, however, that allele i confers a fitness advantage that increases the expected frequency of allele i by 20%, $s = 0.20$, in proceeding generations. We incorporate the fitness advantage of allele i as follows:

$$\begin{aligned} \Pr(Allele\ i\ is\ lost\ in\ next\ generation) &= \left(1 - \frac{1+s}{N}\right)^N \\ &= \left(1 - \frac{1.20}{6}\right)^6 = 0.262 \end{aligned} \tag{1.2-7}$$

Note that even though we expect a 20% increase in the number of individuals carrying allele i in the next generation, allele i is still lost in roughly one out of every four occasions, which is very comparable to the probability calculated using random genetic drift, or $s = 1.0$, in Equation 1.2-6.

On a related note, the opposite case in which a strongly disadvantageous allele with $s = -0.2$ applied to Equation 1.2-7 yields a probability that the allele is lost equal to 0.424, demonstrating that even a highly antagonistic allele, at least in this elementary example, still has a >50% chance of propagating to the next generation. Consequently, simple chance is far more likely to determine whether an allele, advantageous or detrimental, propagates to further generations, particularly when N is small. The effects of selection, though, become substantially more apparent across larger time scales, as opposed to examining a single generation.

1.3 Behavior of selection over time and across different fitness regimes

To demonstrate the contrasting influences of random genetic drift and selection over different time scales, we can scale the theoretical experiment presented in Figure 1.1-2 from $N = 6$ to $N = 1000$ instead. Accordingly, we will examine an allele initially present in half the population such that $i = 500$ or $x_i = 0.5$. Because of the vastly increased population size, N, we expect that fixation or loss of allele i will take a much longer period of time since variance is inversely proportional to sample size. As such, we will model this Wright-Fisher population example over an extended timeframe of 1500 generations and across different selection coefficients, s, to examine the effect of selection across these varying conditions (**Figure 1.3-1**).

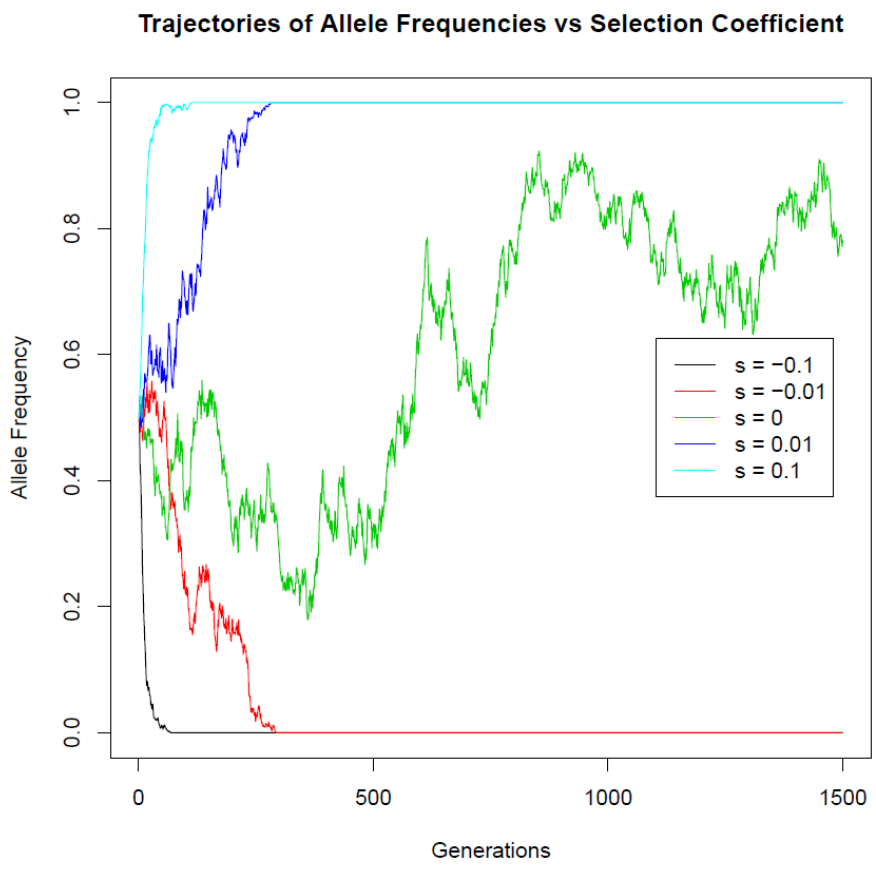


Figure 1.3-1 Impact of selection examined over relatively long time periods. *While the allele under random genetic drift only (green, $s = 0$) approaches fixation, all other alleles under selection are either fixed or lost over much shorter timeframes.*

The extended timeframe in Figure 1.3-1 clearly demonstrates the substantial influence selection has relative to random genetic drift. Even alleles under relatively weaker selection, $s = \pm 0.01$, are fixed or lost in less than 400 generations in this example; however, it is important to remember the length of a human generation in such examples. Under a conservative estimate of 20 years per generation, 400 generations equates to 8,000 years – an extensive time period for an allele to persist. It is also important to note that these selection dynamics are not readily apparent over shorter time frames. For instance, if we take the same experiment setup as in Figure

1.3-1 but instead frame it over 100 generations (or 2000 years if we again assume 20 years per generation), we see that random genetic drift dominates the behavior of allele frequency for alleles not under strong selection (**Figure 1.3-2**).

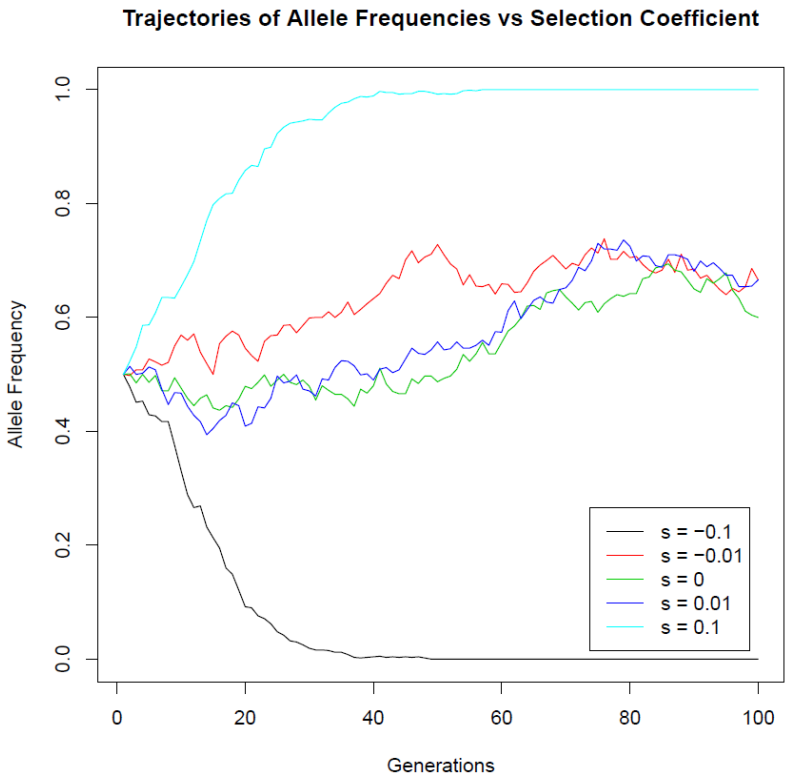


Figure 1.3-2 Impact of selection examined over short time periods. *While alleles under strong selection* $(s \pm 0.1)$ *are quickly fixed or lost, alleles under weak selection* $(s \pm 0.01)$ *behave comparably to alleles under random genetic drift* $(s = 0)$ *over short timeframes.*

The key point from examining the experiments in Figure 1.3-1 and Figure 1.3-2 is that weakly deleterious alleles can persist in populations for a long time period due to random genetic drift. Considering the recent, explosive growth of human populations dating roughly to the agricultural revolution, we expect an influx of relatively young genetic variation segregating at rare and low allele frequencies (Keinan and Clark, 2012). Some fraction of the new variation will be deleterious and should persist in populations as a result of random genetic drift.

All *de novo* variants, though, will start at an allele frequency of $1/N$ as opposed to $0.5N$ as shown in Figures 1.3-1 and 1.3-2. Hence, in addition to calculating the probability that a new allele is lost in a proceeding generation as demonstrated for Figure 1.2-2, we can also calculate the probability that an allele at initial frequency x reaches fixation such that every individual in the population now carries allele i. To do this, we first take note that Equation 1.2-7 is an exponential equation which can be approximated, assuming $N \gg 1$, as follows:

$$\begin{aligned}\Pr(Allele\ i\ is\ lost\ in\ next\ generation) &= \left(1 - \frac{1+s}{N}\right)^N \\ &\approx e^{-(1+s)}\end{aligned} \qquad (1.3\text{-}1)$$

We therefore expect the fixation probability to follow an exponential distribution. The expected fixation probability, $\pi(x)$, is given below (the exact derivation of Equation 1.3-2 requires applying a Taylor series expansion and solving a second order differential equation which is beyond the scope of this document and therefore not presented):

$$\pi(x) = \frac{1 - e^{-2N_e s x}}{1 - e^{-2N_e s}} \qquad (1.3\text{-}2)$$

where x is the initial allele frequency, s is the selection coefficient, and N_e is the effective population size (see Appendix A.1 for more information regarding N_e).

While the fixation probability, $\pi(x)$, derived here is specific to an idealized Wright-Fisher population, Equation 1.3-2 is nonetheless particularly telling when applied to a scenario very relevant to human populations. Specifically, we can ask what shape does the fixation probability take for *de novo* mutations in the population. Setting $x = 1/N$, the frequency of a newly emerged variant in a population, and

applying $N_e = N$, which is applicable for Wright-Fisher populations, into Equation 1.3-2 yields:

$$\pi = \frac{1 - e^{-2s}}{1 - e^{-2Ns}} \tag{1.3-3}$$

We find that the fixation probability, π, for de novo variants depends only on the selection coefficient, s, and the effective population size, N. To gain a firmer understanding of how the selection coefficient, s, impacts the probability of fixation, π, we arbitrarily set $N = 100$ and then plot π with respect to s (**Figure 1.3-3**).

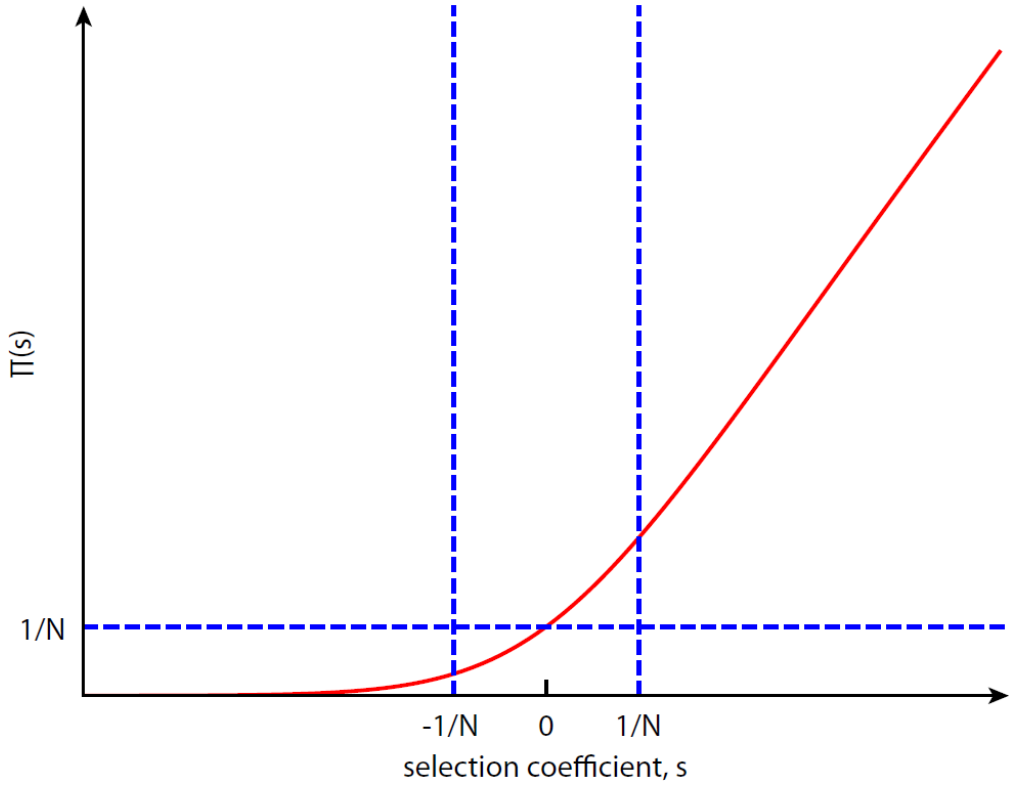


Figure 1.3-3 Plot of Equation 1.3-3 demonstrating how the probability of fixation, π, varies with respect to the selection coefficient, s. *Strongly deleterious, nearly neutral, and strongly advantageous regimes occur at* $s \ll -\frac{1}{N}, -\frac{1}{N} < s < \frac{1}{N}$, *and* $s \gg \frac{1}{N}$, *respectively, and are demarcated by blue vertical lines. Plot is specific to de novo mutations which by definition occur at a frequency of* $x = 1/N$ *and, when neutral, have a the probability of fixation,* $\pi = 1/N$.

Examining Figure 1.3-3, three distinct selection regimes emerge.

1. **Strongly Deleterious Regime** ($s \ll -1/\mathrm{N}$):

   Probability of fixation decreases exponentially as the selection coefficient becomes increasingly negative. Note, though, that the even within this deleterious regime, the probability of fixation is nonzero.

2. **Nearly Neutral Regime** ($-1/N < s < 1/\mathrm{N}$):

   Selection is relatively weak in this regime, and therefore changes in allele frequency for mutations in this regime are heavily influenced by random genetic drift which has a probability of fixation $\pi = 1/N$. It is important to note, though, that the interval in which this nearly neutral regime dominates continuously narrows as the effective population size, represented here by N, increases. Hence, the likelihood that a mutation is neutral decreases as population size increases.

3. **Strongly Advantageous Regime** ($s \gg 1/\mathrm{N}$):

   Probability of fixation increases approximately linearly as the selection coefficient becomes increasingly positive. Nonetheless, a strongly advantageous allele with $s = 0.2$ for $N = 100$ results in a fixation probability of $\pi \approx 0.33$, meaning that in two out of three cases, an advantageous *de novo* mutation with $s = 0.2$ is still lost.

1.4 Concluding remarks on random genetic drift and selection

Explosive population growth in humans has led to a rapid influx of rare variants which are expected to play a role in human disease and complex traits (Keinan and Clark, 2012). To gain a basic insight as towards why we expect an influx of rare, deleterious variants when rapid population growth occurs, we introduced two fundamental

concepts in population genetics: random genetic drift and selection. Using the probability distributions presented in Equations 1.2-6 and 1.3-1, we demonstrated how even a *de novo*, strongly advantageous allele still had an ~26% chance of being lost in the next generation which was only an ~7% decrease in probability versus a strictly neutral allele. Consequently, random genetic drift is heavily influential on the lifecycle of a weakly deleterious allele, as observed in Figure 1.3-3, particularly over relatively short time frames as demonstrated in 1.3-2.

Fitness is defined with respect to the reproduction success rate of an individual with a particular genetic composition. A functional allele that impacts fitness relatively minimally can propagate readily by random genetic drift. Fitness also varies with respect to a changing environment. A trait that is advantageous in one time period or geographic region may be deleterious under a contrasting time period or environment. Therefore, the selection coefficient value for a particular allele is not static but instead is dynamic. Moreover, while selection acts robustly to deplete deleterious alleles, the probability of deleterious alleles fixing is not zero (**Figure 1.3-3**). Such alleles are therefore expected to be present in growing populations where selection has not had sufficient time to effectively purge deleterious *de novo* mutations. Identifying and characterizing such alleles can therefore be very informative towards understanding the impact that rare, recently emerged variation has on human health and traits.

1.5 Human interactome networks and disease

Disease-associated mutations are found to be enriched within non-synonymous

mutation categories, particularly among missense mutations (**Figure 1.5-1**) (Stenson et al., 2014) and are also reported to be enriched within coding regions of the human genome (Hindorff et al., 2009). This enrichment strongly implies that disease-associated mutations frequently induce a local structural perturbation that specifically alters protein function, as opposed to elimination protein function entirely.

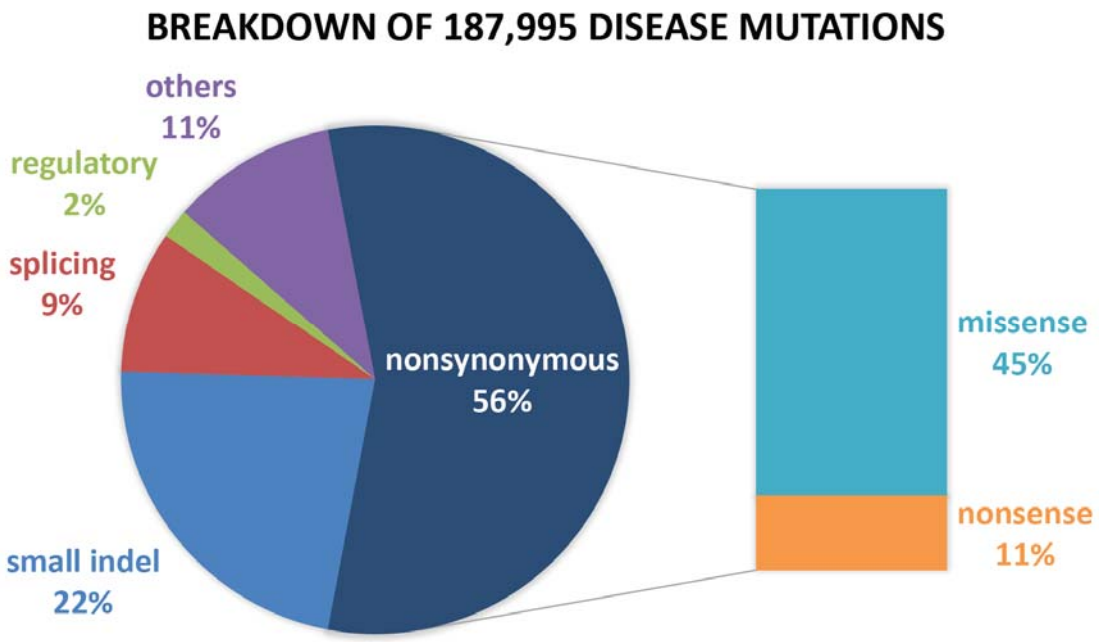


Figure 1.5-1 Human disease-associated mutations categorized in HGMD by mutation type *(accessed August 1, 2016). Note that missense mutations occupy 45% of all known disease-associated mutations.*

The structural position of a mutated amino acid on a protein can provide crucial insight towards the potential impact of a missense mutation in disease. Using structurally-resolved protein-protein interaction networks, researchers found that disease-associated mutations are enriched upon interaction interfaces (Das et al., 2014b; Wang et al., 2012), providing direct evidence that disease-associated mutations frequently function by disrupting specific protein interactions as opposed to

destabilizing protein folding all together. Indeed, a recent study explored the impact of 2,332 disease-associated mutations on protein stability using a LUMIER assay (Sahni et al., 2015). In this assay, destabilized mutant proteins will bind more strongly to protein chaperones, including HSP90 and HSC70. Antibodies specific to particular chaperones are then used to pull down destabilized proteins in complex with specific chaperones (Taipale et al., 2012; Taipale et al., 2014). Using LUMIER, the authors found that 28% of disease-associated mutations bound to at least one tested chaperone. While the authors noted that limited assay sensitivity could suggest that a higher fraction of mutated proteins may be bound to chaperones, nonetheless, 72% of their tested alleles did not strongly destabilize protein folding (Sahni et al., 2015). Considering the enrichment of disease-associated mutations on protein interaction interfaces and the limited impact of disease mutations on protein stability, the authors argue that disease-associated mutations must function primarily through perturbations of specific protein interactions.

The effects of in-frame mutations on protein-protein interactions fall into three categories: (i) non-disruptive, which leave all protein interactions intact; (ii) partially disruptive, which perturb only a subset of protein interactions; and (iii) null-like, which perturb all interaction partners (**Figure 1.5-2**). Mutations have been previously categorized according to this schematic using high-throughput protein-protein interaction assays (Das et al., 2014a; Sahni et al., 2015; Wei et al., 2014; Zhong et al., 2009). While numerous assays for detecting protein-protein interactions are available, including PCA, MAPPIT, and LUMIER (Das et al., 2014a), yeast two-hybrid (Y2H) demonstrates a unique advantage in regards to detecting disrupted protein interactions;

over 16,000 human protein-protein interactions have been detected by Y2H using the Y8800 and Y8930 strains of *Saccharomyces cerevisiae* (Rolland et al., 2014; Rual et al., 2005; Venkatesan et al., 2009; Yu et al., 2011). Using this resource, researchers can directly compare the effect of missense mutations on protein interactions in comparison to retested wild-type interactions. In this manner, researchers have shown that disease-associated mutations primarily result in partially disruptive interaction perturbation profiles (Sahni et al., 2015; Wei et al., 2014; Zhong et al., 2009).

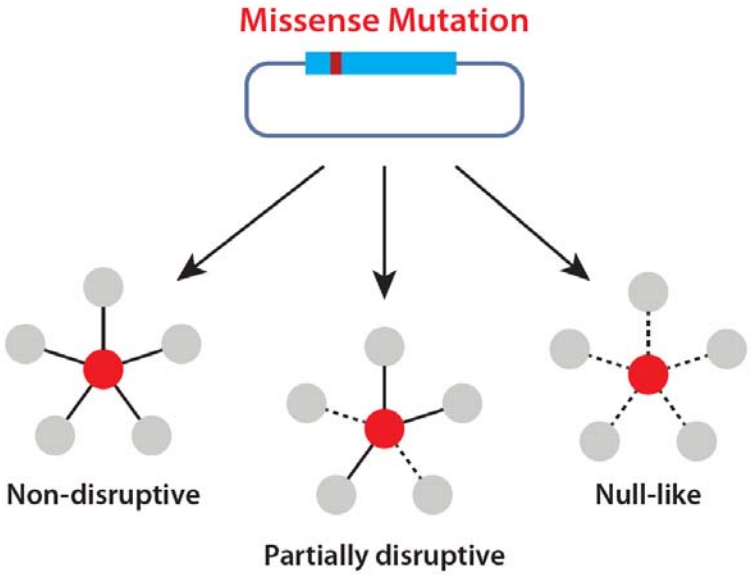


Figure 1.5-2 Missense mutations can have three principle impacts on protein-protein interactions. *Categorizing missense mutations in this manner can provide insight towards the molecular mechanisms of trait- and disease-associated mutations.*

In addition to providing the largest reference database of known human protein-protein interactions, the sensitivity of Y2H is also comparable to that of other high-throughput techniques for detecting protein-protein interactions. Researchers demonstrated this by designing a Positive Reference Set (PRS) of 184 well-established protein-protein interactions that have been reported by multiple publications through

traditional lab techniques such as co-immunoprecipitation. A Random Reference Set (RRS) of 184 completely random proteins pairs, and therefore extremely unlikely to interact, were also included. The authors then compared the PRS and RRS interaction detection rates for Y2H in comparison to other methods including PCA and MAPPIT and found that each method tested had a PRS detection rate (true positive rate, TPR) of ~20% (Braun et al., 2009; Venkatesan et al., 2009). In contrast, while most methods had an RRS detection rate (false positive rate, FPR) of ≥ 5%, the RRS for Y2H was ~1%. The low FPR of Y2H is largely attributed to screening for and removing known autoactivators which can trigger Y2H reporter activity independent of an interaction partner (Walhout and Vidal, 2001). Comparable methods for minimizing false positive rates in other high-throughput techniques have not been readily developed, though a new PCA construct using an infrared reporter protein has shown promise in low-throughput applications (Tchekanda et al., 2014).

Perturbations to protein-protein interaction networks have been conceptualized as frameworks for interpreting complex traits and disease (Vidal et al., 2011; Zhong et al., 2009). In an interactome framework, disruptions in specific modules or to particular edges may manifest as specific organism phenotypes. Such frameworks have even been extended beyond straightforward two-dimensional representations of gene-encoded protein interactions to three-dimensional networks in which the interfaces between interaction interfaces are structurally-resolved (Das et al., 2014b; Meyer et al., 2018; Meyer et al., 2013; Wang et al., 2012).

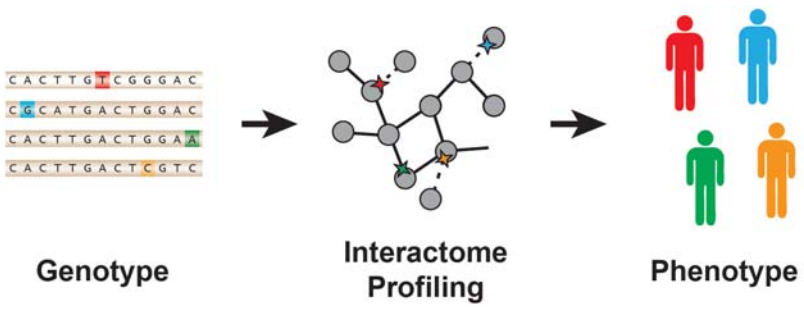


Figure 1.5-3 *Interpreting genotype-to-phenotype relationships through perturbations within a protein-protein interaction network.*

Network-based, structurally-resolved representations of interactome networks are particularly appealing because they also provide a framework from which to interpret non-Mendelian genotype-to-phenotype connections, particularly for cases of locus heterogeneity, in which any mutation in multiple different genes can result in the same disease phenotype, or gene pleiotropy, in which mutations in a single gene map to multiple clinically distinct disorders. Using a structurally resolved protein-protein interaction network, Wang and colleagues identified three disease-associated mutations on the protein WASP that occurred on two distinct interaction interfaces. Two mutations on the interaction interface between WASP and VASP resulted in X-linked thrombocytopenia while a disease-associated mutation on a distinct interaction interfaces that mediated the WASP-CDC42 interaction resulted in X-linked neutropenia, a clinically distinct disease (Wang et al., 2012). Considering that rare variants have been found to be enriched on protein-protein interaction interfaces and have been demonstrated to result in unique interaction perturbation profiles (Khurana et al., 2013), protein-protein interaction networks represent a promising resource for identifying and characterizing function population variants.

Profiling interaction perturbations on a large scale, however, is resource-

intensive and is unlikely to keep pace with current scales of variant discovery. As such, researchers have developed several genome-wide tools for identifying potentially deleterious human polymorphisms, including PolyPhen-2 (Adzhubei et al., 2010), SIFT (Ng and Henikoff, 2003), and CADD (Kircher et al., 2014). While these tools have been expertly developed and rely on fundamental principles of protein conservation and evolution, these tools have performed poorly in clinical settings and seldom result in measurable phenotypes. Indeed, a recent study of 33 *de novo* missense variants in mice found that only 20% of mutations predicted to be deleterious by PolyPhen-2 resulted in discernible phenotypes in mice homozygous for the *de novo* mutations tested (Miosge et al., 2015). A more recent, substantially larger analysis in mice consisting of genotype and phenotype data for 116,330 ENU-induced mutations found that 17% of mutations scored as "probably damaging" by PolyPhen-2 resulted in a discernible phenotypes in mice (Wang et al., 2018). Hence, substantial progress is still needed before functional prediction algorithms alone can be used to reliably detect functional polymorphisms in sequencing data.

1.6 Brief outline of recent high profile large-scale sequencing studies

A long-standing goal of medical genetics is to identify genomic variants that contribute to human disease and complex traits. With this goal in mind, researchers developed the Common Disease – Common Variant hypothesis which proposes that relatively common genetic variants of considerably low penetrance (penetrance is the probability that the carrier of the variant will express the disease) are the major contributors to commonly occurring human disease (Schork et al., 2009). In support of

this hypothesis, numerous Genome-Wide Association Studies (GWAS) have been performed, uncovering hundreds of genetic variants associated with complex human disease and traits (Manolio et al., 2009). Unfortunately, most of the variants identified confer little risk for the associated trait or disease. For example, 18 genetic loci have been linked to Type 2 diabetes; however, the 18 loci only account for 6% of the proportion of heritability of Type 2 diabetes (Zeggini et al., 2008). Since GWAS-identified loci often poorly explain phenotypic variance despite tens of thousands of individuals studied, researchers have proposed alternative explanations for connecting genetic loci to common disease. Specifically, the Rare Variant – Common Disease hypothesis in which rarer variants of larger effect sizes are proposed to account for much of the missing heritability unexplained by GWAS loci (Manolio et al., 2009). Hence, large-scale sequencing efforts of ever-expanding size, including the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015), Exome Sequencing Project (ESP6500) (Fu et al., 2013; Tennessen et al., 2012), UK10K (The UK10K Consortium, 2015), and Exome Aggregation Consortium (ExAC) of 60,706 individual exomes (Lek et al., 2016), have been developed with the goal of identifying putatively functional genetic variation segregating at rare allele frequencies. A summary of these and related studies is provided in Table 1.6-1.

Table 1.6-1 Summary of findings from recent high profile, large-scale sequencing studies

Project	Study Size	Major Conclusions	Citation(s)
1000 Genomes Project	1,092 individual genomes in Phase I; 2,504 individual genomes in Phase III	Covers >99% of SNPs with MAF > 1%. Affirms that African genomes harbor greatest number of variant sites and asserts that 153-320 rare and low-frequency non-synonymous variants per genomes are deleterious.	1000 Genomes Project Consortium, *Nature* 2012 (Phase I), 1000 Genomes Project Consortium, *Nature* 2015 (Phase III)
Exome Sequencing Project (ESP6500)	2,440 individual exomes in Phase I; 6,515 individual exomes in Phase II	Asserts that 2.3% of 13,595 SNVs per individual across 313 genes are predicted to affect protein function and that 95.7% of these functional variants are rare. Attributes excess of rare, deleterious variants to explosive population growth in humans.	Tennessen et al., *Science* 2012 (Phase I), Fu et al., *Nature* 2015 (Phase II)
UK10K	2,440 individual exomes in Phase I; 6,515 individual exomes in Phase II	Performed WGS and WES for nearly 10,000 individuals in UK, uncovering 24 million variants not in European segments of 1000 Genomes or ESP6500. Single marker association tests for 3,781 individuals across 64 phenotypes unveiled no evidence for low-frequency alleles with large effects on traits.	UK10K Consortium, *Nature* 2015
Exome Aggrecation Consortium (ExAC)	60,706 individual exomes	Shows evidence for widespread mutation recurrence with one variant every eight bases. Developed widely-used pLI metric for identifying essential, haploinsufficient genes. Largest reference database of human coding variation to date.	Lek et al., *Nature* 2016
Genomes of the Netherlands (GoNL)	250 family trios or quads (769 individual genomes)	Trio design of study enabled large-scale discovery of *de novo* variants. Asserts that ~63 *de novo* mutations occur per offspring and that the rate *de novo* mutation occurrence increases 2.5% per year of father's age.	Francioli et al., *Nature Genet.* 2014
Genome of Icelanders	2,636 individual genomes	Found an excess of homozygosity and rare coding variation among 20 million SNPs sequenced. Performed a GWAS to identify a frameshift mutation c.234delC (MAF = 0.64%) in MYL4 that caused early-onset atrial fibrillation (OR = 110.3).	Gudbjartsson et al., *Nature Genet.* 2015
Simons Simplex Complex (SSC) for families with ASD	2,517 family trios or quads (~1900 unaffected siblings)	States that 43% of proband *de novo* LGD mutations, in comparison to 13% of proband missense mutations, contribute to ASD and nearly all were opposite WT alleles.	Iossifov et al., *Nature* 2014

1.7 Widespread impact of population variants across global populations

In principle, the rarest variant that can be identified in a single study is defined as having an allele count equal to one divided by the total number of alleles sequenced in the study, otherwise referred to as a singleton. Since humans are diploid, a singleton in a sequencing study of 1000 individuals will have an allele frequency of 1/2000 = 0.05%. Hence, increasingly larger study sizes are needed to find rarer alleles, a challenging task. Rapidly improving next-generation sequencing platforms, though, have readily enabled whole-genome and whole-exome sequencing studies at the large scales needed to probe rare variants. The first major iteration towards this effort was referred to as the 1000 Genomes Project. Split across multiple phases, Phase I of the 1000 Genomes Project aimed to provide a geographical and functional spectrum of allelic variation across 1,092 individuals, representing 14 different populations (The 1000 Genomes Project Consortium, 2012). This effort generated 38 million unique SNPs and covered 98% of "accessible" (for reference, variants that reside within regions of the genome that are difficult to sequence, such as homopolymeric regions, are considered inaccessible) SNPs with minor allele frequency (MAF) > 1.0%, in addition to unprecedented coverage for rare alleles.

Despite uncovering 38 million unique SNPs, only a fraction of these are expected to be functionally relevant. In an effort to identify such variants, researchers focused on conserved genomic sites, defined as having a GERP (Cooper et al., 2005) score > 2.0, in which the GERP score is a metric that uses multiple sequence alignments to identify highly constrained genomic sites, and then examined the distribution of both synonymous and non-synonymous variants across the site

frequency spectrum (SFS). In this manner, the researchers were able to identify excess rare variants sufficiently deleterious such that they would never reach common alleles frequencies (MAF > 5.0%). Accordingly, the researchers estimated, on average, that every individual carries 153 to 320 deleterious coding variants in their genomes. In addition to these SFS-identified deleterious variants, the average genome also carries 150 Loss-of-Function variants (e.g. stop-gain and splice-site mutations) as well as 20-40 clinically relevant mutations, determined by their expected presence in the Human Gene Mutation Database, HGMD (Stenson et al., 2009).

Phase III of the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015) expanded their whole-genome study size from 1,092 individuals to 2,504 individuals across 26 populations and uncovered a total of 88 million variants. The study further estimated that an average 4.1 to 5.0 million sequence differences are expected between two genomes and noted that African genomes harbor the greatest number of variant sites. This is in agreement with the Out of Africa hypothesis since more recently arisen populations such as Europeans have smaller and more recent founder populations which, in turn, lowers nucleotide diversity. Lastly, while the study notes that 64 million of the variants discovered were rare (MAF < 0.5%), the majority of variants per individual genome are actually common. Only 40,000 to 200,000 of the 4.1 to 5 million variants per genome (1-4%) are actually rare. This is a subtle but important distinction. Databases including those for the 1000 Genomes Project, ESP6500, and ExAC consistently report the total number of unique variants found in any individual. Common variants (MAF > 5.0%) are therefore counted only once despite occurring in at least one out of every twenty genomes. In contrast, consider a

singleton which by definition can only occur in one individual. Each singleton is therefore guaranteed to increase the variant count in a database which is not true of common variants. As a result, every large-scale sequencing database will assert that rare variants dominate their studies. This is certainly true. However, despite an enrichment for rare variants in human genomes, it is worth reiterating that the vast majority of variants per individual genome, >96% for the 1000 Genomes Project, are not rare. This is a crucial distinction to keep in mind when surveying literature on sequencing projects since that means that even if a common variant has only a mildly deleterious effect, that effect will be widespread which cannot be said for rare variants.

Noting that disease-associated variants are enriched within coding regions of the genome (Hindorff et al., 2009) and capitalizing on the lowered expense of sequencing whole exomes as opposed to whole-genomes, sample sizes increased again for the ESP6500 publications. In Phase I of ESP6500, 2,440 individual exomes from European and African American (EA and AA, respectively) ancestries were sequenced and ~500,000 single nucleotide variants (SNVs) were uncovered, 86% of which were rare (MAF < 0.5%) (Tennessen et al., 2012). (See Appendix A.3 for a short discussion regarding the difference between the terms SNPs and SNVs) The authors further remark that explosive population growth in combination with weak purifying selection accounts for the excess of rare variation observed in their study. More importantly, the authors also report that out of an average of 13,595 SNVs per individual, 318 to 580 (2.3-4.2%) SNVs are predicted to impact protein function. The authors arrive at this estimate in the same manner in which the authors from the Phase I iteration of the 1000 Genomes Project performed their genome-wide estimate for deleterious

variation: specifically, the distribution of both synonymous and non-synonymous variants across the site frequency spectrum is examined to identify excess rare variants sufficiently deleterious that they would never reach common alleles frequencies. Note that this 318 to 580 figure reported in ESP6500 Phase I is markedly higher than the 153 to 320 deleterious coding variants reported in the 1000 Genomes Project despite using the same method to identify deleterious variation (The 1000 Genomes Project Consortium, 2012). This suggests that increased sample size may account for the higher deleterious variant counts reported in ESP Phase I.

Purifying selection on ESP variants were also measured using seven functional prediction algorithms, including widely used tools such as PolyPhen-2 (Adzhubei et al., 2010) and SIFT (Ng and Henikoff, 2003). While each method attempts to find likely deleterious alleles at conserved genomic sites, individual predictions can vary widely between methods. Indeed, the authors report that 47% of nonsynonymous variants are predicted to be functional by at least one method while only 1% of nonsynonymous variants are predicted to be functional by all seven methods tested in their publication (Tennessen et al., 2012). Because of this limited agreement, the authors, perhaps arbitrarily, adopt a "majority rule" for assigning deleterious variation in which at least four out of the seven methods must agree that a variant is functional for the variant to be considered deleterious. While the majority rule indicates that 16.9% of nonsynonymous SNVs are functional, no evidence is presented supporting the proposition that prediction methods yield more accurate results when combined with other prediction methods. Moreover, the authors never justify how the 16.9% of deleterious variants identified by their majority rule method differs with respect to the

2.3 to 4.2% of variants per genome proposed to alter protein function using their site frequency spectrum-deleterious approach. Presumably, the methods differ in that both common and rare variation can be scored as deleterious using functional prediction algorithms while the SFS-based method for identifying deleterious, medically-relevant variants is limited to rare variants. Considering, though, that functional prediction algorithms like PolyPhen-2 are heavily biased towards identifying rare variants as damaging, it is still uncertain why two different methods for classifying deleterious variation were used by the authors.

ESP6500 Phase II expanded their exome sequencing set to 6,515 individuals primarily of EA and AA ancestry, uncovering a total of >1 million SNVs (Fu et al., 2013). To differentiate their study from their previous work beyond increased sample size (Tennessen et al., 2012), the authors instead focused on estimating the allele age for each of their sequenced SNVs with the specific goal of investigating how recent explosive population growth has yielded an excess of weakly deleterious SNVs segregating at rare allele frequencies. Notably, the authors report that 73% of all coding SNVs arose in the past 5,000 to 10,000 years, in concurrence with subsequent explosive human population growth. Similar to their previous work (Tennessen et al., 2012), the authors again apply a "majority rule" to identify putatively functional SNVs that are deemed as potentially deleterious by at least four separate functional prediction algorithms and report that 86% of these putatively functional alleles arose within this same 5,000 to 10,000 year timeframe. In other words, the vast majority of mutation burden imposed on human genomes is due to deleterious SNVs that arose as a consequence of recent explosive population growth.

The authors also show proof in support of the Out of Africa hypothesis by performing simulations similar in theme to those presented for Figure 1.3-1 and 1.3-2 (but executed with significantly more sophistication) to estimate the probability that *de novo* weakly deleterious SNVs at differing selection coefficients and arising at different times in the past, could still survive to this day. In comparison to neutral alleles, $s = 0$, the authors found that deleterious alleles with selection coefficients $0 \geq s \geq -0.1\%$ survived at ~50% or more, depending on the magnitude of s, of the rate at which neutral alleles survived over a 50,000 year time period. (For clarity, the authors are not stating that 50% of all deleterious alleles survive. Recall that under a Wright-Fisher model, a *de novo* neutral allele at frequency $1/N$ has a fixation probability equal to $1/N$. The authors are therefore stating that weakly deleterious alleles survive in the population at a rate comparable to the rate at which a *de novo* neutral allele would survive within relatively short time frame.) Extending beyond 50,000 years lowers the survival rate of deleterious alleles with $s = 0.1\%$ to nearly 0% such that most weakly deleterious SNVs are purged at time frames beyond 100,000 years. Similarly, the proportion of deleterious variants with an estimated age of 50,000 to 100,000 years in disease-associated genes was enriched in EAs in comparison to AAs, in agreement with the Out of Africa model.

Lastly, but importantly, the authors noted that a rare coding SNV was observed every 1 out of 52 and 1 out of 57 base pairs in EAs and AAs, respectively; however, the rate at which rare variation was found varied profoundly with respect to the disease-relevant KEGG pathways to which a gene belonged. In general, significantly younger alleles (and therefore higher rates of rare variation per base pair) were

observed in disease-relevant genes in comparison to genes from less essential pathways. This observation appears to agree with a previous report in which 202 disease-relevant, drug-related genes were sequenced in 14,002 individuals (Nelson et al., 2012). In their study, Nelson and colleagues report that a rare variant (MAF < 0.5%) occurs every 1 out of 17 base pairs in these essential genes and assert that the elevated rate of rare variation observed is a consequence of recent explosive population growth in humans. Similar to the ESP Phase I study (Tennessen et al., 2012), Nelson and colleagues also report that ~20% of all mutations in their 202 drug-related genes are expected to be deleterious using a variety of functional prediction algorithms.

Realizing how valuable reference databases for human genetic variation are towards the medical and functional interpretation of variants, the Exome Aggregation Consortium (ExAC) was constructed as a resource composed of coding sequencing data from 14 exome sequencing cohorts (Lek et al., 2016). The combined resource represents over 10 million unique variants from 60,706 individuals, representing seven population groups. The unprecedented size of ExAC unveiled widespread mutational recurrence, resulting in an average of one variant per every eight coding bases. This unprecedented sequencing scale enabled Lek and colleagues to take an in-depth measure of the functional constraints acting on each gene by examining the fraction of missing variation within a gene compared to neutral expectation. Their basic notion was that decreasingly less sequencing variation should be observed across increasingly essential genes.

Using Z-scores to quantify deviations in expected variant counts from

expectation, the authors found that missense variants and protein-truncating variants (PTVs) in particular were depleted with respect to neutral expectation while synonymous variants showed no such depletion. In agreement with conclusions from other large-scale sequencing studies (Fu et al., 2013; The 1000 Genomes Project Consortium, 2015; The UK10K Consortium, 2015), this depletion for nonsynonymous variation implies that, as a whole, purifying selection limits the prevalence of nonsynonymous variation in human genomes. Since PTVs are under stronger functional constraint than missense variants, and more likely to result in a deleterious loss-of-function (LoF) phenotype as a result, the authors then measured observed and expected PTV counts per gene and then grouped all analyzed genes into three categories: (1) presumed null genes, in which observed PTV counts were roughly equal to expected PTV counts; (2) recessive genes, in which observed PTV counts were less than 50% of expected PTV counts; and (3) haploinsufficient genes, in which observed PTV counts were 10% or less of expected PTV counts. The authors then optimized these metrics using an expectation-maximizing algorithm to construct their pLI metric, the probability that a gene is intolerant to loss-of-function variants. The authors found that pLI-categorized LoF-intolerant genes (pLI ≥ 0.9) overlapped with virtually all known haploinsufficient human disease genes and, at its most constrained, overlapped well with genes involved in fundamental biological processes, including those in the spliceosome, ribosome, and proteasome. Lastly, the authors reported that the near tenfold increase in size relative to ESP6500 allowed for more effective filtering of potential disease-associated mutations with implausibly high allele frequencies. To this end, the authors report that the average exome in ExAC contains

54 variants likely incorrectly listed as disease-associated in databases such as HGMD or ClinVar due to their high MAFs.

Currently, ExAC serves as the largest reference database for human variation available, but as new whole-genome and whole-exome sequencing efforts come to completion, an even larger repository sequencing database known as gnomAD representing 123,136 exomes and 15,496 genomes is soon to supplant ExAC.

1.8 Identifying genotype-to-phenotype associations via large-scale sequencing

While the 1000 Genomes Project, ESP6500, and ExAC sequencing studies have effectively constructed reference panels for human genetic variation spanning global populations, population-specific sequencing efforts focused on genomes from Dutch (The Genome of the Netherlands Consortium, 2014), Icelandic (Gudbjartsson et al., 2015), and UK populations (The UK10K Consortium, 2015) have also surfaced. These high profile studies have, unsurprisingly at this point, unveiled extensive rare coding variation specific to these population subgroups but, unlike their predecessor studies, also provide examples of rare variants linked to specific disease phenotypes. For example, a whole-genome sequencing study of 2,636 Icelanders was incorporated into a GWAS for early-onset atrial fibrillation on a subset of 1,294 individuals with hospital discharge diagnoses (Gudbjartsson et al., 2015). This study revealed a rare frameshift mutation (MAF = 0.64%) in the myosin light chain gene *MYL4* which had an extremely convincing odds ratio = 110.3. This *MYL4* frameshift mutation, c.234delC, was homozygous in eight individuals, all of which were diagnosed with early-onset atrial fibrillation. Similar connections were reported for rare mutations in

ABCB4 in association with liver diseases, highlighting the power of rare variant association tests performed in large population cohorts.

Rare variant associations with common and complex disease were also reported in a larger study of ~10,000 individuals from the UK10K project (The UK10K Consortium, 2015). In an effort to explore the genetic components of quantitative traits, the UK10K study included phenotype data for 64 traits, including obesity, diabetes, and blood pressure. Single marker association tests were performed on a subset of 3,781 individuals with available whole-genome sequencing data. 27 genetic loci associations with 31 traits were reported, only two of which, though, were found to be novel. The two loci in genes *ADIPOQ* and *APOC3*, which are involved in regulating glucose and triglyceride levels, respectively, occurred at relatively low frequencies but did not have large effect sizes. Similarly, the authors also identified three rare LoF variants in the lipoprotein gene *APOB* of varying effect size. As a whole, however, association statistics appeared underpowered towards detecting true signal. As a result, the authors found no evidence of low-frequency alleles with large effects upon traits. Moreover, no linked loci were reported when examining complex traits such as obesity, autism, or schizophrenia.

Of the many large-scale sequencing studies performed in European populations, the Genomes of the Netherlands (GoNL) study particularly stands out since it was performed specifically on 250 family trios or quads. As such, the GoNL study allowed for functional characterization of *de novo* variants, which are spontaneous mutations found in offspring and therefore not transmitted to them by their parents. *De novo* mutations should not be confused with singletons. Both occur at

an allele frequency of one over the total number of alleles sequenced; however, singletons cannot be verified as spontaneous as opposed to inherited without sequencing the parents of the individual carrying the singleton as well.

The GoNL authors report that in addition to a vast influx of previously undiscovered variants segregating at rare allele frequencies, offspring genomes on average carry 63 *de novo* mutations. Furthermore, the rate of *de novo* mutation occurrence is reported to increases by 2.5% per year of the father's age. This is a particularly interesting result because *de novo* mutations are under no prior functional constraint and are therefore strong candidates for study in human disease, particular in family trios in which the child is afflicted with a disease while the parents or siblings are unaffected. While no disease phenotype data is presented in the GoNL study, the authors still report that the average individual carries 60 LoF variants, ~56 of which occur at common allele frequencies and therefore assumed to be benign. Similarly, 484 missense variants are predicted to be damaging, of which 93% are common.

In stark contrast to the GoNL study, a particularly high profile study on the impact of *de novo* variation in complex disease was performed using the Simons Simplex Collection (SSC) dataset (Iossifov et al., 2014). The SSC consists of whole-genome sequencing data for ~2,500 families consisting of parent-offspring trios or quads with two unaffected parent, an ASD-afflicted proband and, for ~1900 sequenced families, an unaffected sibling (Fischbach and Lord, 2010). Due to this trio and quad design, *de novo* variants specific to ASD phenotypes can be identified.

Iossifov and colleagues examined the distribution of *de novo* variants in ASD-afflicted probands and unaffected siblings and found that nearly all proband *de novo*

coding variants occurred opposite wild-type alleles, suggesting that haploinsufficient gene are principally targeted (Iossifov et al., 2014). The authors also note that the *de novo* synonymous mutation rate was relatively equal between ASD probands and unaffected siblings at 0.34 and 0.33 mutations per child, respectively. In contrast, likely gene-damaging (LGD, e.g. nonsense and splice-site mutations) *de novo* mutations occurred at rate of 0.21 and 0.12 mutation per child in ASD probands and unaffected siblings, respectively. By subtracting this "ascertainment differential" between ASD probands and unaffected siblings (0.21 – 0.12) and dividing this differential by the mutations rate per child in ASD probands (0.09 ÷ 0.21), the authors determined that 43% of proband LGD events contribute to ASD. Furthermore, the authors reported that only 13% of proband missense events contribute to ASD but the confidence interval was wide, so the authors focused their analysis on LGD mutations instead.

By grouping genes targeted by LGD mutations in ASD probands in comparison to unaffected siblings, Iossifov and colleagues go on to report that genes involved in fundamental developmental pathways, including genes regulated by the FMRP transcription factor, are targeted by proband mutations. Moreover, ASD-afflicted children with LGD mutations in FMRP-regulated genes had, as a whole, a 20 point drop in their nonverbal IQ scores. Correspondence between age of parent and ASD risk was also reported. However, it is imperative to note that only 391 *de novo* LGD mutations in ASD probands were found. Consequently, in ~2100 ASD-afflicted probands, a *de novo* LGD mutation does not contribute to ASD diagnosis. Considering that 1500 *de novo* missense mutations were also identified, it stands to reason that a

substantial fraction of functional missense variation must be contributing to ASD outcomes. Identifying such functional variation, though, remains a significant challenge, especially considering that the authors do not report any results for missense mutations predicted to be damaging.

1.9 Concluding remarks

Common themes arise among the 1000 Genomes (The 1000 Genomes Project Consortium, 2012, 2015) and ESP6500 (Fu et al., 2013; Tennessen et al., 2012) manuscripts. Each gives evidence to widespread rare genetic variation as a result of recent, explosive population growth. Applying functional prediction algorithms to their datasets suggest that ~20% of segregating coding variants are potentially deleterious and that deleterious variation is enriched among rare variants, a result also found in the GoNL study (The Genome of the Netherlands Consortium, 2014) and Nelson and colleagues' study in 202 drug-targeted genes (Nelson et al., 2012). More constrained, site frequency spectrum-based predictions, though, conclude that up 4.2% of nonsynonymous variants per individual genome may disrupt protein function, though no mechanisms or direct evidence of altered protein function is ever presented. Indeed, recent studies have suggested that the functional prediction methods used to assert widespread functional coding mutations rarely result in measureable phenotypes, at least in mice (Miosge et al., 2015; Wang et al., 2018). Nonetheless, the influx and enrichment of rare, widespread genetic variation invariably has an impact on human health and disease (Keinan and Clark, 2012). Pinpointing just how and which variants mechanistically impact human phenotypes at either an organism or

molecular level, though, has proven difficult. Interactome-based approaches, however, have proven resourceful towards dissecting different mechanisms of disease-associated mutation and could prove equally insightful when searching for potentially functional population variants from ever-emerging sequencing data.

1.10 References

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat Meth *7*, 248-249.

Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J.M., Murray, R.R., Roncari, L., de Smet, A.-S.*, et al.* (2009). An experimentally derived confidence score for binary protein-protein interactions. Nat Meth *6*, 91-97.

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

Das, J., Fragoza, R., Lee, H.R., Cordero, N.A., Guo, Y., Meyer, M.J., Vo, T.V., Wang, X., and Yu, H. (2014a). Exploring mechanisms of human disease through structurally resolved protein interactome networks. Molecular BioSystems *10*, 9-17.

Das, J., Lee, H.R., Sagar, A., Fragoza, R., Liang, J., Wei, X., Wang, X., Mort, M., Stenson, P.D., Cooper, D.N.*, et al.* (2014b). Elucidating Common Structural Features of Human Pathogenic Variations Using Large-Scale Atomic-Resolution Protein Networks. Human Mutation *35*, 585-593.

Fischbach, G.D., and Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron *68*, 192-195.

Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Altshuler, D., Shendure, J., Nickerson, D.A.*, et al.* (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216-220.

Gudbjartsson, D.F., Helgason, H., Gudjonsson, S.A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B.V., Hjartarson, E.*, et al.* (2015). Large-scale whole-genome sequencing of the Icelandic population. Nature Genetics *47*, 435.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-

wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences *106*, 9362-9367.

Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K., Vives, L., Patterson, K.E.*, et al.* (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216-221.

Keinan, A., and Clark, A.G. (2012). Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. Science *336*, 740-743.

Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A.*, et al.* (2013). Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. Science *342*.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet *46*, 310-315.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B.*, et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285-291.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A.*, et al.* (2009). Finding the missing heritability of complex diseases. Nature *461*, 747.

Meyer, M.J., Beltrán, J.F., Liang, S., Fragoza, R., Rumack, A., Liang, J., Wei, X., and Yu, H. (2018). Interactome INSIDER: a structural interactome browser for genomic studies. Nature Methods *15*, 107.

Meyer, M.J., Das, J., Wang, X., and Yu, H. (2013). INstruct: a database of high-quality 3D structurally resolved protein interactome networks. Bioinformatics.

Miosge, L.A., Field, M.A., Sontani, Y., Cho, V., Johnson, S., Palkova, A., Balakishnan, B., Liang, R., Zhang, Y., Lyon, S.*, et al.* (2015). Comparison of predicted and actual consequences of missense mutations. Proceedings of the National Academy of Sciences *112*, E5189-E5198.

Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St. Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.-A., Fraser, D.*, et al.* (2012). An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. Science *337*, 100-104.

Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res *31*, 3812-3814.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, Samuel J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R.*, et al.* (2014). A Proteome-Scale Map of the Human Interactome Network. Cell *159*, 1212-1226.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N.*, et al.* (2005). Towards a proteome-scale map of the human protein-protein interaction network. Nature *437*, 1173-1178.

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J.I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G.I., Wang, Y.*, et al.* (2015). Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders. Cell *161*, 647-660.

Schork, N.J., Murray, S.S., Frazer, K.A., and Topol, E.J. (2009). Common vs. Rare Allele Hypotheses for Complex Diseases. Current opinion in genetics & development *19*, 212-219.

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

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

Taipale, M., Krykbaeva, I., Koeva, M., Kayatekin, C., Westover, K.D., Karras, G.I., and Lindquist, S. (2012). Quantitative Analysis of Hsp90-Client Interactions Reveals Principles of Substrate Recognition. Cell *150*, 987-1001.

Taipale, M., Tucker, G., Peng, J., Krykbaeva, I., Lin, Z.-Y., Larsen, B., Choi, H., Berger, B., Gingras, A.-C., and Lindquist, S. (2014). A quantitative chaperone interaction network reveals the architecture of cellular protein homeostasis pathways. Cell *158*, 434-448.

Tchekanda, E., Sivanesan, D., and Michnick, S.W. (2014). An infrared reporter to detect spatiotemporal dynamics of protein-protein interactions. Nature Methods *11*, 641.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G.*, et al.* (2012). Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science *337*, 64-69.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68.

The Genome of the Netherlands Consortium (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet *46*, 818-825.

The UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82-90.

Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.-I.*, et al.* (2009). An empirical framework for binary interactome mapping. Nat Meth *6*, 83-90.

Vidal, M., Cusick, Michael E., and Barabási, A.-L. (2011). Interactome Networks and Human Disease. Cell *144*, 986-998.

Walhout, A.J.M., and Vidal, M. (2001). High-Throughput Yeast Two-Hybrid Assays for Large-Scale Protein Interaction Mapping. Methods *24*, 297-306.

Wang, T., Bu, C.H., Hildebrand, S., Jia, G., Siggs, O.M., Lyon, S., Pratt, D., Scott, L., Russell, J., Ludwig, S.*, et al.* (2018). Probability of phenotypically detectable protein damage by ENU-induced mutations in the Mutagenetix database. Nature Communications *9*, 441.

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotech *30*, 159-164.

Wei, X., Das, J., Fragoza, R., Liang, J., Bastos de Oliveira, F.M., Lee, H.R., Wang, X., Mort, M., Stenson, P.D., Cooper, D.N.*, et al.* (2014). A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. PLoS Genet *10*, e1004819.

Yu, H., Tardivo, L., Tam, S., Weiner, E., Gebreab, F., Fan, C., Svrzikapa, N., Hirozane-Kishikawa, T., Rietman, E., Yang, X.*, et al.* (2011). Next-generation sequencing to generate interactome datasets. Nat Meth *8*, 478-480.

Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I.W., Abecasis, G.R., Almgren, P., Andersen, G.*, et al.* (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet *40*, 638-645.

Zhong, Q., Simonis, N., Li, Q.-R., Charloteaux, B., Heuze, F., Klitgord, N., Tam, S., Yu, H., Venkatesan, K., Mou, D.*, et al.* (2009). Edgetic perturbation models of human inherited disorders. Mol Syst Biol *5*.

CHAPTER 2

A MASSIVELY PARALLEL PIPELINE TO CLONE DNA VARIANTS AND EXAMINE MOLECULAR PHENOTYPES OF HUMAN DISEASE MUTATIONS

2.1 Preface

Chapter 2 was originally published as "Wei, X.*, Das, J.*, Fragoza, R.*, Liang, J.*, Bastos de Oliveira, F.M., Lee, H.R., Wang, X., Mort, M., Stenson, P.D., Cooper, D.N., Lipkin, S.M., Smolka, M.B., Yu, H. A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. *PLoS Genet* **10**, e1004819 (2014)" where * indicates co-first author. Per guidelines by the Field of Biochemistry, Molecular, and Cellular Biology, the manuscript has been amended to focus primarily on contributions made by Robert Fragoza but has retained the results of others for the purposes of clarity when reading the text.

2.2 Abstract

Understanding the functional relevance of DNA variants is essential for all exome and genome sequencing projects. However, current mutagenesis cloning protocols require Sanger sequencing, and thus are prohibitively costly and labor-intensive. We describe a massively-parallel site-directed mutagenesis approach, "Clone-seq", leveraging next-generation sequencing to rapidly and cost-effectively generate a large number of mutant alleles. Using Clone-seq, we further develop a comparative interactome-scanning pipeline integrating high-throughput GFP, yeast two-hybrid (Y2H), and mass spectrometry assays to systematically evaluate the functional impact of mutations on

protein stability and interactions. We use this pipeline to show that disease mutations on protein-protein interaction interfaces are significantly more likely than those away from interfaces to disrupt corresponding interactions. We also find that mutation pairs with similar molecular phenotypes in terms of both protein stability and interactions are significantly more likely to cause the same disease than those with different molecular phenotypes, validating the *in vivo* biological relevance of our high-throughput GFP and Y2H assays, and indicating that both assays can be used to determine candidate disease mutations in the future. The general scheme of our experimental pipeline can be readily expanded to other types of interactome-mapping methods to comprehensively evaluate the functional relevance of all DNA variants, including those in non-coding regions.

2.3 Author Summary

With rapid advances in sequencing technologies, tens of millions of DNA variants have now been discovered in the human population. However, there are currently no experimental methods available for examining the impact of DNA variants in a high-throughput fashion. As a result, we have no functional data on the vast majority of these variants, which is a major roadblock to generating novel biological insights and developing new disease prevention therapeutic strategies. To address this issue, we have successfully developed the first massively-parallel site-directed mutagenesis approach, Clone-seq, to leverage the power of next-generation sequencing to generate a large number of mutant alleles in a fast and cost-effective manner. In conjunction with Clone-seq, we established a high-throughput comparative interactome-scanning

pipeline to experimentally elucidate the effect of variants on protein stability and interactions. Additionally, Clone-seq can be used to generate clones for all DNA variants, including those in non-coding regions.

2.4 Introduction

Owing to rapid advances in next-generation sequencing technologies, tens of thousands of disease-associated mutations (Stenson et al., 2009) and millions of single nucleotide polymorphisms (SNPs) (Fu et al., 2013; The 1000 Genomes Project Consortium, 2012) have been identified in the human population. With the large number of ongoing whole-exome and whole-genome sequencing projects (Fu et al., 2013; The 1000 Genomes Project Consortium, 2012) hundreds of thousands of new SNPs are now being discovered every month. Hence, there is an urgent need to develop high-throughput methods to sift through this deluge of sequence data and rapidly determine the functional relevance of each variant. Here, we focus on coding variants, firstly because trait- and disease-associated SNPs are significantly over-represented in nonsynonymous sites (Hindorff et al., 2009), and secondly because the vast majority of disease-associated mutations identified to date reside within coding regions (Stenson et al., 2009). We evaluate the functional impact of coding variants by examining their effects on corresponding protein-protein interactions, because most proteins carry out their functions by interacting with other proteins (Vidal et al., 2011).

Recent studies have begun to use large-scale protein interaction networks to understand human diseases and their associated mutations (Vidal et al., 2011; Zhong et al., 2009). By integrating structural details with high-quality protein networks, we

created a 3D interactome network where the interface for each interaction has been structurally resolved (Wang et al., 2012). Using this 3D network, we demonstrated that in-frame disease mutations (missense mutations and in-frame insertions/deletions) are significantly enriched at the interaction interfaces of the corresponding proteins (Wang et al., 2012). Our results indicate that alteration of specific interactions is very important for the pathogenesis of many disease genes, highlighting the importance of 3D structural models of protein interactions in understanding the functional relevance of coding variants. However, many important questions still remain unanswered – for example, what fraction of protein-protein interactions is altered by disease mutations to cause the corresponding disorders? Furthermore, do structural details of the interacting proteins, especially the position of the mutation relative to the interaction interface, affect the ability of a given disease mutation to alter a specific interaction?

To address these questions, we decided to focus on proteins with known disease mutations that participate in interactions with available co-crystal structures in the Protein Data Bank (PDB) (Berman et al., 2000). To detect the alteration of the interactions by disease mutations, it is necessary to first detect the interactions of the wild-type proteins using an assay of choice. This turns out to be a major bottleneck because all high-throughput interaction-detection assays have very limited sensitivity (Braun et al., 2009; Yu et al., 2008). Our assay of choice is Y2H because there are over 16,000 human protein interactions detected by our version of Y2H that can serve as the reference interactome for comparison (Rolland et al., 2014; Rual et al., 2005; Venkatesan et al., 2009; Yu et al., 2011), the largest for any assay performed to date (**Appendix B.1**).

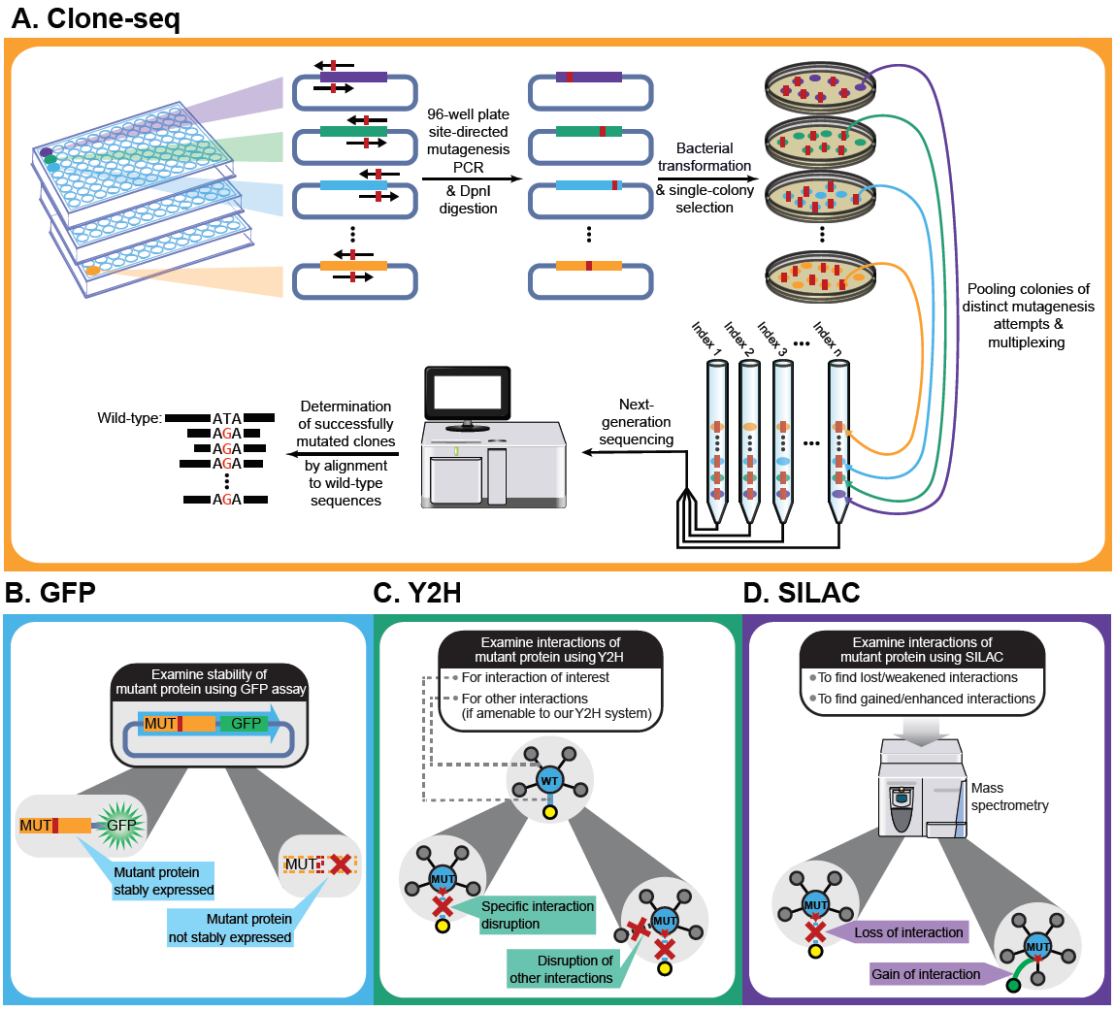


***Figure 2.4-1* Schematic of our comparative interactome-scanning pipeline.** Our pipeline begins with Clone-seq (a), a massively-parallel low-cost site-directed mutagenesis pipeline leveraging next-generation sequencing. This is followed by a high-throughput GFP assay (b) to determine protein stability, and a high-throughput Y2H assay (c), along with SILAC-based mass spectrometry (d) to determine the impact of DNA coding variants on protein interactions.

In total, there are 217 interactions detected by our version of Y2H with available co-crystal structures; 51 of these also have known missense disease mutations on corresponding proteins in the Human Gene Mutation Database (HGMD) (Stenson et al., 2009) and the corresponding interactions for the wild-type proteins can be detected

in our experiments with strong Y2H-positive phenotypes (**Append B.2**; **Materials and Methods**). Here, we focused on missense mutations because they are intrinsically more likely to generate interaction-specific disruptions (Zhong et al., 2009). We established a high-throughput comparative interactome-scanning pipeline to clone disease mutations and examine their molecular phenotypes (**Figure 2-4.1**). The methodologies established here can be readily applied to any non-synonymous variant in the coding region, including nonsense mutations.

2.5 Results

2.5.1 Clone-seq: A massively parallel site-directed mutagenesis pipeline using next-generation sequencing

The first step of our pipeline is a massively parallel approach, termed Clone-seq, designed to leverage the power of next-generation sequencing to generate a large number of mutant alleles using site-directed mutagenesis in a rapid and cost-effective manner. Current protocols for site-directed mutagenesis require picking individual colonies and sequencing each colony using Sanger sequencing to identify the correct clone (Suzuki et al., 2005). This standard approach is both labor-intensive and expensive; therefore, it does not scale up to genome-wide surveys. In Clone-seq, we put one colony of each mutagenesis attempt into one pool (**Figure 2.4-1a**; in other words, each pool contains one and only one colony for each desired mutation) and combine multiple pools through multiplexing for one Illumina sequencing run (Salehi-Ashtiani et al., 2008). Colonies for generating different mutations of the same gene can be put into the same pool, which can be easily distinguished computationally

when processing the sequencing results. This is true even for mutations occurring at the same site (**Figure 2.5-1a**).

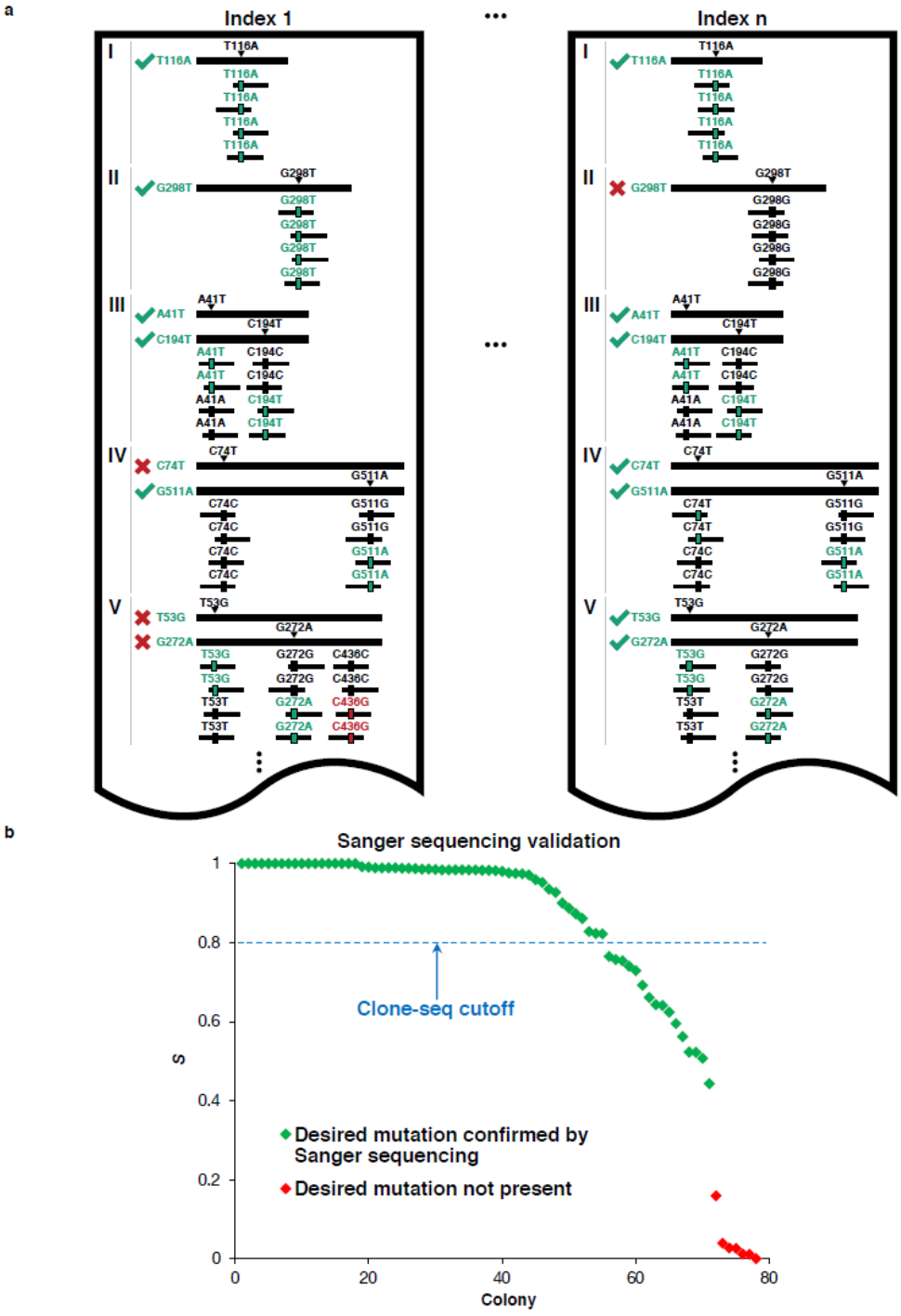


Figure 2.5-1 Identifying usable clones from Clone-seq. (a) Schematic illustrating criteria used to determine which of the clones generated by our Clone-seq pipeline are usable for

further assays – green ticks indicate usable clones, while red crosses indicate clones that cannot be used. (b) Variation of *S* across different mutagenesis attempts that either contain or do not contain the desired mutation as confirmed by Sanger sequencing.

For the 51 selected interactions, we chose 27 disease-associated mutations of residues at the interface ("interface residue"), 100 mutations in the rest of the interface domain ("interface domain") and 77 mutations away from the interface ("away from the interface"; **Figure 2.5-2a,b**). These interfaces were determined using solvent accessible surface area calculations as previously described (Das et al., 2014; Khurana et al., 2013) on 7,340 co-crystal structures (**Materials and Methods**). To set up our Clone-seq pipeline, we first started with 39 mutations from these 204 and picked 4 colonies for each mutation. As a reference, we also pooled together all the wild-type alleles in our human ORFeome library to be sequenced together with the 4 pools of the mutagenesis colonies. In total, there were 40.1 million Illumina HiSeq 1×100 bp reads for our Clone-seq samples (**Appendix B.3.1**) for an average of >2,500× coverage on all desired mutation sites. Therefore, our Clone-seq pipeline has the capacity to generate >3,000 mutations in one full lane of a HiSeq run with 1×100 bp reads, drastically improving the throughput and decreasing overall sequencing costs by at least 10-fold (**Appendix B.3.3**).

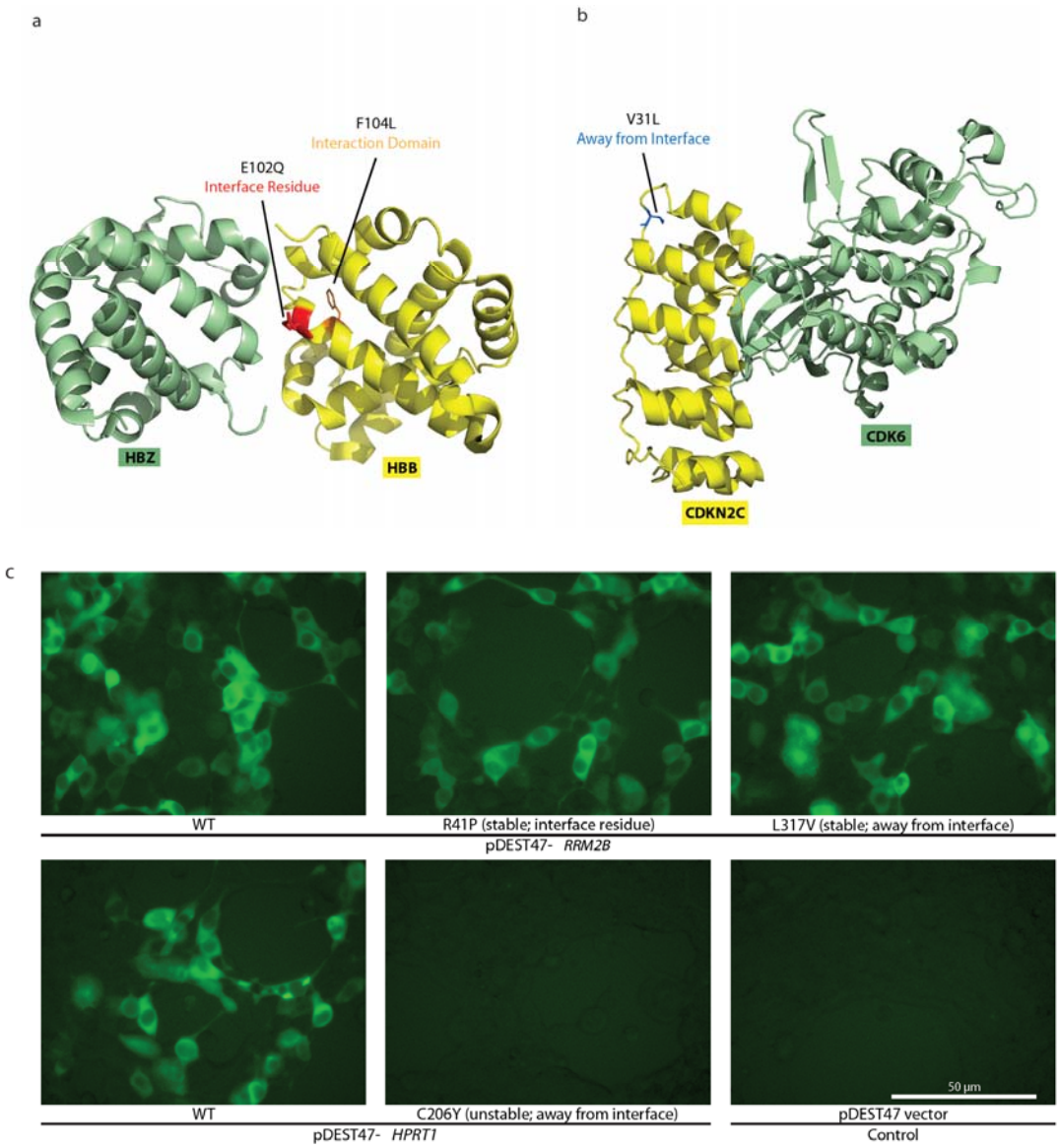


***Figure 2.5-2** Examples of disease mutations in different structural loci of protein-protein interactions and examples of our GFP assay results.* *(a) Crystal structure (PDB id: 3W4U) depicting a D100Y mutation (on Hbb) at an interface residue and a F104L mutation in the interface domain for the Hbb-Hbz interaction. (b) Crystal structure (PDB id: 1G3N) depicting a V31L mutation (on Cdkn2c) away from the Cdkn2c-Cdk6 interaction interface. (c) GFP assays that determine the stability of wild-type Rrm2b and the R41P and L317V mutations on Rrm2b that are at an interface residue and away from the interface for the Rrm2b-Rrm2b interaction; GFP assays that determine the stability of wild-type Hprt1 and the C206Y mutation on Hprt1 that is away from the interaction interface of Hprt-Hprt1. Empty vector was used as a negative control.*

Figure 2.5-1a presents a schematic of the criteria we use to determine which clones contain the desired mutation and can be used for subsequent steps. For example, in pool 1, all reads (ignoring sequencing errors) confirm that genes I and II each contain the desired mutation – T116A and G298T, respectively. For gene III, we want to generate two separate clones with two separate mutations – III_{A41T} and III_{C194T}. Since half the reads contain T41 (instead of A41) and the other half contain T194 (instead of C194), and we normalize DNA concentrations across all samples, we can infer that both mutant clones were generated successfully. In contrast, for gene IV, we see that while half the reads contain A511 (instead of G511), all the reads are wild-type at C74. Thus, we infer that while the IV_{G511A} clone is successfully generated, the IV_{C74T}clone is not. For gene V, although both mutant clones are successfully generated, half the reads contain an additional mutation, C436G. Since it is impossible to know which of the two clones for V contains this unwanted mutation, neither clone is usable. Similarly, we can determine mutant clones I_{T116A}, III_{A41T}, III_{C194T}, IV_{C74T}, IV_{G511A}, V_{T53G}, and V_{G272A} as usable clones in pool *n*. Based on these criteria, we developed the *S* score calculation and used it to determine successful mutagenesis attempts (**Materials and Methods**). Out of 156 colonies for 39 mutations, 125 of them contain the desired mutations (S>0.8), an overall 80% PCR-mutagenesis success rate. In fact, we were able to pick correct clones for all 39 mutant alleles using only the first two pools in Clone-seq. All 78 clones from the first two pools, from which the correct ones were selected for use in subsequent steps, were also Sanger sequenced for verification. 55 Clone-seq positive results with S>0.8 were all confirmed and there is a

clear separation in the *S* scores between the successful and failed mutagenesis attempts (**Figure 2.5-1b**).

One major advantage of our Clone-seq pipeline is that it allows us to carefully examine whether other unwanted mutations have been inadvertently introduced during PCR-mutagenesis in comparison with the corresponding wild-type alleles, since we obtain reads spanning the entire gene. We found that there are on average 4–5 unwanted mutations introduced in each pool of 39 colonies. This corresponds to a 0.013% PCR error rate (**Materials and Methods**), in agreement with previous studies (Vandenbroucke et al., 2011). The detection of unwanted mutations, especially those distant from the mutation of interest, is achieved in traditional site-directed mutagenesis pipelines by Sanger sequencing through the gene of interest. This is costly and labor-intensive, especially because multiple sequencing runs are needed for one long gene. However, since Clone-seq yields reads spanning the entire gene, we were able to determine which of the generated clones definitely do not have unwanted mutations in the full length of their sequences as illustrated in Figure 2.5-1a (**Materials and Methods**), and we pick only these clones for subsequent assays.

To further test our Clone-seq pipeline, we applied it to generate clones for 113 SNPs on 66 genes from the recently published Exome Sequencing Project dataset (Fu et al., 2013). Using the same approach as described above, we sequenced 4 colonies each for the 113 alleles of interest using one third of a 1×100 bp MiSeq run. We obtained 4.7 million reads for these 113 alleles. With a threshold of $S>0.8$, we were able to determine that 370 out of the 452 colonies (82%) contain the desired mutation, in perfect agreement with the PCR-mutagenesis success rate obtained earlier. We were

able to choose colonies that contain only the desired mutation for all 113 alleles. Because the whole MiSeq run produced 17.7 million reads and we only used 4.7 million for generating the 113 mutant clones, the capacity of our Clone-seq pipeline using one full lane of a 1×100 bp HiSeq run is estimated to be >3,000, exactly the same as our previous assessment (**Appendix B.3.1)**.

Overall, our pipeline has been significantly optimized to make it very efficient. We established a web tool (http://www.yulab.org/Supp/MutPrimer) to design mutagenesis primers both individually and in batch. MutPrimer can design ~1,000 primers for ~500 mutations in one batch in less than one second. All of the 2,068 primers for the 1,034 mutations in this study were generated by MutPrimer. All mutagenesis PCRs are performed in batch using automatic 96-well procedures. Since single colony picking after bacterial transformation of mutagenesis PCR product is a rate-limiting step, we rigorously optimized this step and found that adding 10 µL mutagenesis PCR products to 100 µL competent cells and plating 50 µL transformed cells give the best transformation yield and well-separated single colonies. Furthermore, rather than individually streaking transformed cells onto agar plates one sample at a time, we were able to significantly increase throughput by spreading colonies using glass beads onto four sector agar plates which are partitioned into four non-contacting quadrants (**Materials and Methods**). In this manner, a 96-well plate of transformed bacteria can be plated out onto 24 four-sector agar plates in ~15 minutes. Traditional site-directed mutagenesis pipelines require miniprepping each of the selected colonies and sequencing them separately by Sanger sequencing. To drastically improve the throughput of our Clone-seq pipeline, we pooled together the

bacteria stock of a single colony for each mutagenesis attempt to perform one single maxiprep, which makes the library construction step much more efficient and amenable to high-throughput. Furthermore, existing variant calling pipelines (McKenna et al., 2010) cannot be applied to our Clone-seq results because the expected allelic ratios built into these pipelines are a function of the ploidy of the organism. However, in our Clone-seq pipeline there is no concept of ploidy. We pool together many mutations for one gene in the same pool (e.g., 40 mutations for *MLH1*) and different genes often have different numbers of mutations. Our *S* score calculation and unwanted mutation detection pipeline was designed according to our pooling strategy (**Materials and Methods**).

In total, we have used the novel Clone-seq pipeline successfully to generate 1,034 (39+113+882) mutant clones without any additional unwanted mutations, confirming the scalability, accuracy, and throughput of our Clone-seq pipeline.

2.5.2 A high-throughput GFP assay to determine the impact of mutations on protein stability

For the 204 mutations on proteins with co-crystal structures, we first examined whether the mutant proteins can be stably expressed in human cells. To do this, we tagged every wild-type and mutant protein with GFP at the C-terminus using high-throughput Gateway cloning (**Figure 2.4-1b**). The GFP constructs were transfected into HEK293T cells and fluorescence intensities were measured by a plate reader (**Figure 2.5-2c**; **Materials and Methods**). All fluorescence intensity readings were also confirmed manually under a microscope.

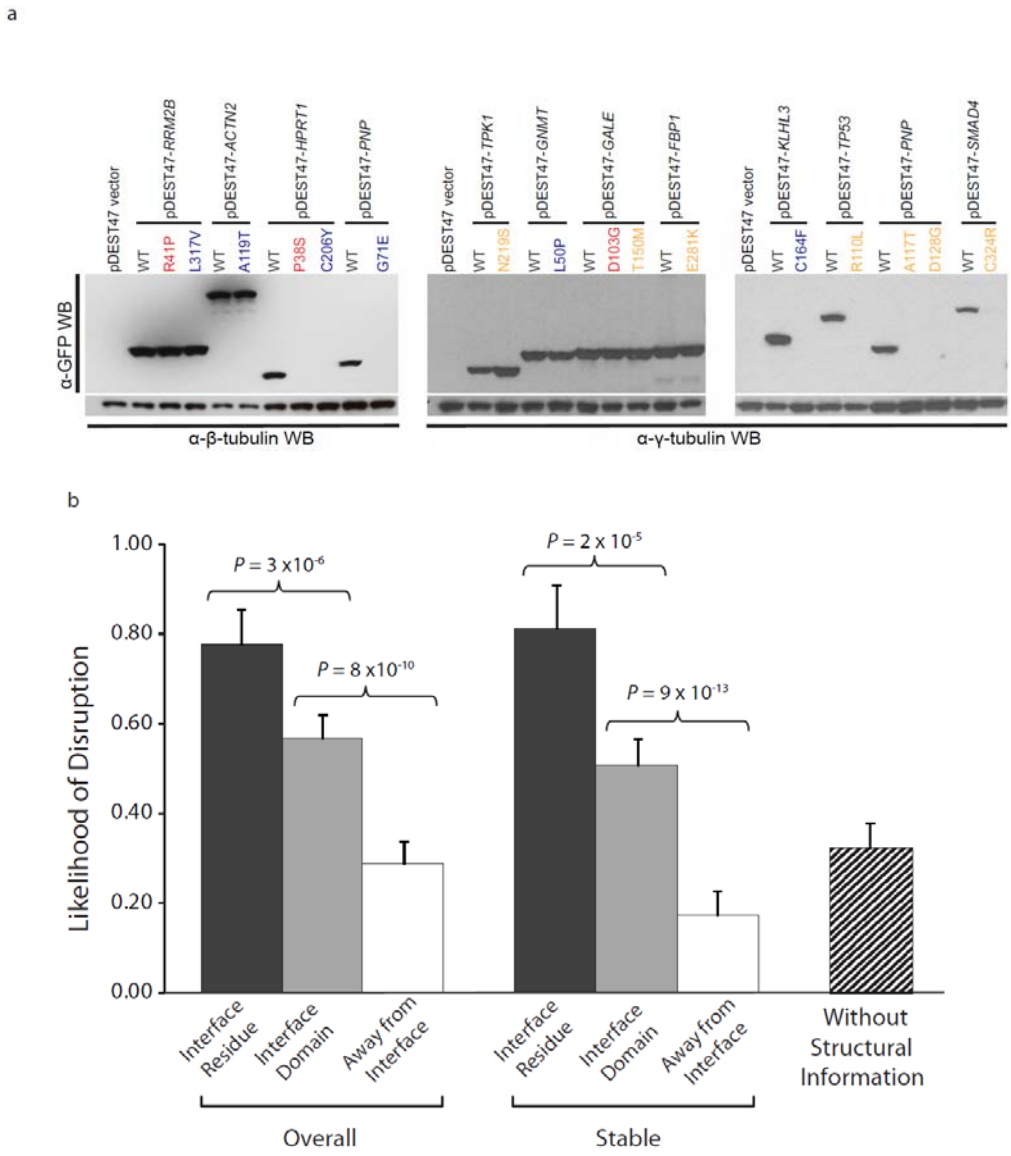


***Figure 2.5-3** Effect of disease mutations on protein stability and protein-protein interactions.* *(a) Western blotting with anti-GFP antibody confirming the protein expression levels of wild-type Rrm2b, Actn2, Hprt1, Pnp, Tpk1, Gnmt, Gale, Fbp1, Klhl3, Tp53, Pnp, Smad4, and corresponding mutant alleles. β-tubulin and γ-tubulin were used as loading controls. Red denotes "interface residue" mutations, orange denotes "interface domain" mutations and blue denotes "away from the interface" mutations. (b) Likelihood of disruption of interactions by "interface residue", "interface domain" and "away from the interface" mutations – overall and for stable mutants only; likelihood of a disease mutation disrupting a given interaction in the absence of structural information. Error bars indicate +SE. (N=204 mutations).*

Compared with the corresponding wild-type proteins, the expression levels of 3 of the 27 "interface residue" mutants, 8 of the 99 "interface domain" mutants and 6 of the 77 "away from the interface" mutants are significantly diminished (**Figure 2.5-2c**; **Materials and Methods**). To validate these findings, we also performed Western blotting for 8 random mutants that are stably expressed and 8 random mutants with significantly diminished expression levels (**Figure 2.5-3a**). Western blotting results confirm our GFP intensity readings.

2.5.3 A high-throughput Y2H assay to determine the impact of mutations on protein interactions

Next, we investigated whether these mutations could affect protein-protein interactions using Y2H (**Figure 2.4-1c**; **Materials and Methods**). We found that 21 of the 27 (78%) "interface residue" mutations, 57 of the 100 (57%) "interface domain" mutations, and only 22 of the 77 (29%) "away from the interface" mutations disrupt the corresponding interactions, thereby demonstrating a clear difference (**Figure 2.5-3b**; $P=3\times10^{-6}$ between "interface residue" and "interface domain" and $P=8\times10^{-10}$ between "interface domain" and "away from the interface") in terms of ability to interfere with protein-protein interactions between mutations at different structural loci within the same protein. Furthermore, comparing with the GFP results, we found that all destabilizing mutations were shown to disrupt the corresponding interactions in our Y2H experiments. By considering only the mutations that do not affect protein expression based on the GFP experiments, we found the same difference: 13 out of 18 (72%) "interface residue" stable mutations, 42 out of 83 (51%) "interface domain"

stable mutations, and only 9 out of 52 (17%) "away from the interface" stable mutations disrupt the corresponding interactions (**Figure 2.5-3b;** $P=2\times10^{-5}$ between "interface residue" and "interface domain" and $P=9\times10^{-13}$ between "interface domain" and "away from the interface"). Since these interfaces are obtained from actual co-crystal structures, our results suggest that accurate structural information can help determine the functional impact of mutations on protein-protein interactions. Wild-type proteins corresponding to 113 of the 153 stably expressed mutant proteins also interact with other proteins as determined by our Y2H experiments (114 interactions in total, termed "other interactions"); however, for these interactions, there are currently no co-crystal structures available in the PDB. Using these other interactions, we calculated the likelihood of a given mutation disrupting a specific interaction without any structural information to be 32% (**Figure 2.5-3b**).

2.5.4 Relationships between measured molecular phenotypes and corresponding disease phenotypes

We then analyzed whether the molecular phenotypes measured by our high-throughput GFP and Y2H assays are correlated with corresponding disease phenotypes. We first examined how mutation pairs on the same gene affect protein stability and its relationship to their corresponding diseases. We find that pairs of mutations that are either both stable or both unstable cause the same disease in 68% and 70% of cases, respectively. However, pairs comprising one stable and one unstable mutation cause the same disease in only 30% of cases ($P=6\times10^{-9}$ and 8×10^{-10}, respectively, **Figure 2.5-4a**). For example, we find that the mutations R727C

and L844F on the spindle checkpoint kinase Bub1b both cause the protein to become unstable and lose all its interactors.

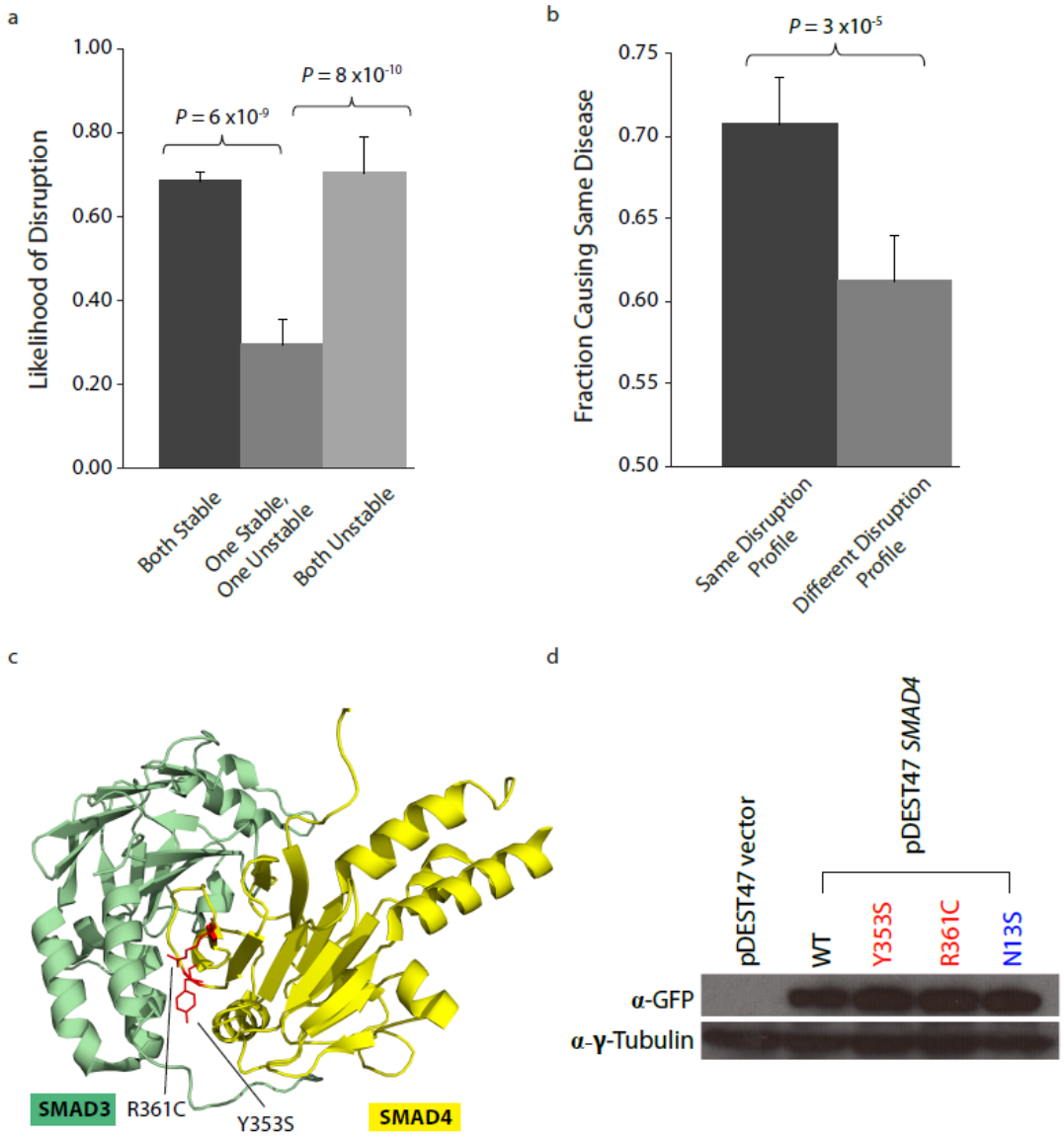


Figure 2.5-4 Relationships between molecular phenotypes and disease phenotypes. *(a) Fraction of mutation pairs on the same gene that cause the same disease: for the same and different effects on protein stability. (b) Fraction of mutation pairs on the same gene that cause the same disease: for the same and different interaction disruption profiles. Error bars indicate +SE. (c) Crystal structure (PDB id: 1U7F) depicting the Y353S and R361C mutations (on Smad4) at interface residues for the Smad4-Smad3 interaction. (d) Y2H analysis of the*

effects of Smad Y353S, R361, and N13S mutations on its interactions with Smad3, Lmo4, Rassf5, and Smad9. Western blotting with anti-GFP antibody confirming the protein expression levels of wild-type Smad4 and its 3 mutant alleles – Y353S, R361C and N13S. γ-tubulin was used as a loading control.

These mutations are both associated with the same disease, mosaic variegated aneuploidy, an autosomal recessive disorder that causes predominantly trisomies and monosomies of different chromosomes (Hanks et al., 2004; Suijkerbuijk et al., 2010). Since our GFP assay shows that these two mutations cause loss of protein product, our results are consistent with Matusuura et al.'s finding that a more than 50% decrease in Bub1b activity leads to abnormal mitotic spindle checkpoint function and mosaic variegated aneuploidy (Matsuura et al., 2006).

We then examined whether mutation pairs on the same gene disrupt the same set or different sets of interactions (i.e., their interaction disruption profiles) and investigated whether their disruption profiles correlates with disease phenotypes. We found that mutation pairs with the exact same disruption profile are significantly more likely to cause the same disease than those with different profiles (70% and 61% respectively, $P=3\times10^{-5}$, **Figure 2.5-4b**). For example, we found that two mutations on Smad4, R361C and Y353S, disrupt its interactions with Smad3 and Smad9 while leaving the interactions with Lmo4 and Rassf5 unaltered (**Figure 2.5-4c**). These two mutations both cause juvenile polyposis coli (Houlston et al., 1998; Roth et al., 1999), a disease is known to be caused by disruption of the core Smad/Bmp signaling pathways (Massagué, 2008). Our Y2H results clearly demonstrate that the R361C and Y353S mutations disrupt the Smad4-Smad3 and Smad4-Smad9 interactions (**Figure 2.5-4c**) leading to disruption of core Smad signaling pathways. However, the mutation

N13S on Smad4 does not disrupt any of these interactions (**Figure 2.5-4c**) and is associated with a different disease, pulmonary arterial hypertension. Our results agree with Nasim et al.'s finding that the N13S mutation does not alter downstream Smad signaling (Nasim et al., 2011). Our findings provide support for the hypothesis that the N13S mutation either impacts pathways outside the core Smad signaling network or are pathogenic only when combined with other environmental and genetic factors (Machado, 2012).

Overall, these results show that mutation pairs with similar molecular phenotypes in terms of both protein stability and interactions are significantly more likely to cause the same disease than those with different molecular phenotypes. This confirms that the molecular phenotypes measured by our high-throughput GFP and Y2H assays are biologically relevant *in vivo*. Furthermore, by comparing the molecular phenotypes, in particular the protein interaction disruption profiles, of mutations/variants to those of known disease mutations, potential candidate mutations for a variety of diseases can be identified.

2.6 Discussion

We have successfully developed the first massively parallel site-directed mutagenesis pipeline, Clone-seq, using next-generation sequencing. Our Clone-seq pipeline is entirely different from previously described random mutagenesis approaches (Araya et al., 2012; Fowler et al., 2010; Pitt and Ferré-D'Amaré, 2010; Starita et al., 2013). Clone-seq is used to generate a large number of specific mutant clones with desired mutations; each individual mutant clone has a separate stock and different clones can

therefore be used separately for completely different downstream assays. In random mutagenesis, a pool of sequences containing different mutations for one gene is generated using error-prone PCR or error-prone DNA synthesis. Therefore, it is not possible to separate one mutant sequence from another and the whole pool can only be used for the same assay(s) together. Furthermore, it is not possible to control which or how many mutations are generated on each DNA sequence. In fact, to improve coverage, most random mutagenesis pipelines generate on average two or more mutations on each DNA sequence (Fowler et al., 2010), which makes it impossible to distinguish the functional impact of each individual mutation on the same sequence. Site-directed mutagenesis and random mutagenesis are designed for different goals: if one wants to generate all possible mutations for a certain protein without the need to separate different clones, it would be more favorable to use random mutagenesis; whereas if one needs to have separate clones for each mutation, site-directed mutagenesis is required. As a result, the two approaches are complementary and not comparable.

While there are highly efficient methods for random mutagenesis (Araya et al., 2012; Fowler et al., 2010; Pitt and Ferré-D'Amaré, 2010; Starita et al., 2013), current protocols for site-directed mutagenesis are low-throughput and become prohibitively expensive if a large number of clones needs to be generated. Clone-seq directly addresses the necessity for a high-throughput site-directed mutagenesis pipeline. It is a robust, cost-effective and efficient method that can be used to generate a total of ~3,000 distinct mutant clones in one full lane of a 1×100 bp HiSeq run. Clone-seq is suitable both for generating mutations across many genes as well as a large number of

mutations on a few genes. The former situation is applicable when one wants to generate many mutations/variants from large-scale studies (e.g., whole-genome or whole-exome sequencing) since they typically identify mutations/variants on a large number of genes (Stransky et al., 2011; The Cancer Genome Atlas Network, 2012). The latter situation usually arises in a study focused on a single pathway with a few genes of interest (e.g., an alanine-scanning mutagenesis to determine functional sites on a gene of interest (Cunningham and Wells, 1989)).

Integrating with Clone-seq, we also established a comprehensive comparative interactome-scanning pipeline, including high-throughput GFP, Y2H, and mass spectrometry assays, to systematically evaluate the impact of human disease mutations on protein stability and interactions. We examine each mutation individually, rather than looking at their combinatorial effects because these inherited germline disease mutations are extremely rare. Therefore, the probability of having even two of these in the same individual becomes infinitesimally small. Our results reveal that the overall likelihood of a given disease mutation disrupting a specific interaction is 32%. Accurate structural information of these interactions obtained from co-crystal structures greatly improves our understanding of the impact of disease mutations: 13 out of 18 (72%) “interface residue” stable mutations, 42 out of 83 (51%) “interface domain” stable mutations, and only 9 out of 52 (17%) “away from the interface” stable mutations disrupt the corresponding interactions, unveiling a clear dependence of the molecular phenotypes of disease mutations on their structural loci. These estimates are not affected by the false negative rate of our Y2H assay as we only use those interactions for which we can detect the wild-type interaction with strong Y2H

phenotypes. Thus, any observed disruption is due to the mutation of interest and not an assay false negative. Furthermore, our Y2H pipeline has been shown to be of high quality and has an experimentally measured false positive rate of ~5% or lower in different organisms (*Arabidopsis* Interactome Mapping Consortium, 2011; Das et al., 2013; Venkatesan et al., 2009; Yu et al., 2008). In addition, the interactions used to understand the relationship between molecular phenotypes and structural loci of disease mutations are all supported by co-crystal structures, therefore these interactions are not assay false positives. We also find that the molecular phenotypes detected by our GFP and Y2H assays correlate with known disease phenotypes, confirming the *in vivo* biological significance of our measurements.

Our comparative interactome-scanning pipeline described and validated here can be applied to experimentally determine in a high-throughput fashion the impact on protein stability and protein-protein interactions for thousands of DNA coding variants and disease mutations, which can directly lead to hypotheses of concrete molecular mechanisms for follow-up studies. Furthermore, the elucidation of molecular phenotypes of disease mutations is also vital for selecting actionable drug targets and ultimately for making therapeutic decisions. Finally, the general scheme of our pipeline can be readily expanded to other interactome-mapping methods, particularly other protein-protein (Braun et al., 2009), protein-DNA (Berger et al., 2006; Reece-Hoyes et al., 2011), protein-RNA (Yakhnin et al., 2012), and protein-metabolite interaction assays (Bandyopadhyay et al., 2012), to comprehensively evaluate the functional relevance of all DNA variants, including those in non-coding regions.

2.7 Materials and Methods

2.7.1 Selecting interactions with mutations on and away from the interface

To calculate atomic-resolution interaction interfaces, we systematically examined a comprehensive list of 7,340 PDB co-crystal structures. To define the interface, we used a water molecule of diameter 1.4 Å as a probe and calculated the relative solvent accessible surface areas of the interacting pair as well as the individual proteins involved in the interaction. Residues whose relative accessibilities change by more than 1 $Å^2$ are considered as potential interface residues, because amino acids at the interface reside on the surfaces of the corresponding proteins, but will tend to become buried in the co-crystal structure as the two proteins bind to each other (Franzosa and Xia, 2011). So, for these residues, there should be a significant decrease in accessible surface area when we compare the bound and unbound states of the protein chains. To identify interface domains, we required at least one of the following criteria to hold:

1. 3did (Stein et al., 2011) or iPfam (Finn et al., 2005) have identified the domain pair as interacting and each of the interface domains contains at least one interface residue based on our calculations.
2. The domain pair contains 5 or more interface residues for each protein according to our calculations.

We then identified the subset of these interactions that contain at least one disease mutation and are amenable to our version of Y2H (Rolland et al., 2014; Rual et al.,

2005; Venkatesan et al., 2009; Yu et al., 2008). Subsequently, we performed a pairwise retest of all these interactions and selected the ones that yield strong Y2H phenotypes, because subsequent steps involve detecting a significant decrease in these phenotypes.

2.7.2 Primer design for site-directed mutagenesis

Primers for site-directed mutagenesis were selected based on a customized version of the protocol accompanying the Stratagene QuikChange Site-Directed Mutagenesis Kit (200518). The following criteria are used:

1. The primer should be of length 30–50 bp and should contain the mutation of interest in the center or one base away.
2. The GC content of the primer should be ≥40% and the primer should start and end with a G or a C.
3. The T_m for the primer should be ≥78°C. T_m was calculated using the following expression:

$$T_m = 81.5 + 0.41 \times (\%GC) - \frac{675}{N} - \%mismatch$$

where N is the primer length in bases, *%GC* is the percentage of G or C nucleotides in the primer, and *%mismatch* is the percentage of mismatched bases in the primer. Values for *%GC* and *%mismatch* are whole numbers.

For cases where no primer satisfies all three criteria simultaneously, we relaxed criterion 2 to GC content ≥30%. We established a supplementary web tool

(http://www.yulab.org/Supp/MutPrimer) to design mutagenesis primers individually or in bulk.

2.7.3 Construction of mutant alleles using high-throughput site-directed mutagenesis PCR

All wild-type clones were obtained from the human ORFeome v8.1 collection (Yang et al., 2011). To generate mutant alleles, sequence-verified single-colony wild-type clones and their corresponding mutagenic primers were aliquoted into individual wells of 96-well PCR plates. Mutagenesis PCR was then performed as specified by the New England Biolabs (NEB) PCR protocol for Phusion polymerase (M0530L), noting that PCR was limited to 18 cycles. The samples were then digested by *Dpn*I (NEB R0176L) according to the manufacturer's manual. After digestion, samples were transformed into competent *E. coli*. Since single colony picking after bacterial transformation of mutagenesis PCR product is a rate-limiting step, we rigorously optimized this step. First, we tried different volumes of competent cells for transformation and found that single colony yields peak when ~100 µL of competent cells are used. It is also necessary to use ~10 µL of mutagenesis PCR product: any lower volume of PCR product results in significantly reduced colony yields, while higher volumes of PCR product do not increase yield. Finally, colony picking was done using four-sector agar plates (VWR 25384-308) that are partitioned into four non-contacting quadrants with glass beads poured onto each plate quadrant. Each bead-filled quadrant was inoculated with ~50 µL of transformed bacteria. This was then spread by lightly shaking the four-sector agar plate. Our optimized transformation

protocol results in a large number of well-separated single colonies that can be easily picked the next day. Upon recovery, single colonies from each quadrant were then picked and arrayed into 96-deepwell plates filled with 300 µL of antibiotic media. Four colonies per allele were picked for next-generation sequencing.

2.7.4 DNA library preparation for Illumina sequencing

DNA library preparation was performed using NEBNext DNA Library Prep Master Mix Set for Illumina (NEB E6040S) according to the manufacturer's manual. Briefly, 5 µg of pooled plasmid DNA (~100 µL, all samples were normalized to the same concentration) was sonicated to ~200 bp fragments. The fragmented DNA was first mixed with NEBNext End Repair Enzyme for 30 mins at 20°C. Blunt-ended DNA was then incubated with Klenow Fragment for 30 mins at 37°C for dA-Tailing. Subsequently, NEBNext Adaptor was added to dA-Tailed DNA. Adaptor-ligated DNA (~300 bp) was size-selected on a 2% agarose gel. Size-selected DNA was then mixed with one of the NEBNext Multiplex Oligos (NEB E7335S) and Universal PCR primers for PCR enrichment. At each step, DNA was purified using a QIAquick PCR purification kit (Qiagen 28104). Multiplexed DNA samples were combined and analyzed in one lane of a 1×100 bp run by Illumina HiSeq 2500.

2.7.5 Identifying successful instances of site-directed mutagenesis based on next-generation sequencing

The mutant colonies were barcoded and pooled as shown in Figure 2.4-1a. The multiplexed colonies were then run on an Illumina sequencer (2 HiSeq runs and 1

MiSeq run) to give 1×100 bp reads. These reads were then de-multiplexed and mapped to the genes of interest using the BWA "aln" algorithm (Li and Durbin, 2009). For each allele, we identified all reads that mapped to the position of the mutation of interest (R_{all}) and those that actually contained the desired mutation (R_{mut}). We then calculated a normalized score (S) that quantifies the fraction of reads containing the desired mutation:

$$S = \frac{R_{mut}}{\frac{1}{k} R_{all}} = \frac{k \times R_{mut}}{R_{all}}$$

where k is the number of different mutations for the same gene.

For 39 mutations, we Sanger sequenced two mutant colonies per mutagenesis attempt to quantify the correlation between S and observation of the desired mutation. We found that all clones with S>0.44 are confirmed to be correct via Sanger sequencing with a clear separation between those that are correct and those that are not (**Figure 2.5-1b**). However, to further ensure that the clones we picked were correct, we require S>0.8 for a colony to be scored as containing the desired mutation.

2.7.6 Identifying unwanted mutations

One major advantage of our Clone-seq pipeline over traditional site-directed mutagenesis protocols using Sanger sequencing (Suzuki et al., 2005) is that we can now carefully examine whether there are other unwanted mutations inadvertently introduced during the PCR process, in comparison with the corresponding wild-type alleles. It is essential to use clones with no unwanted mutations for downstream

experiments, as the presence of these will make it impossible to determine whether the observed disruption is due to the desired or other undesirable mutation(s).

We use samtools "mpileup" (Li et al., 2009) to obtain read counts for different alleles at each nucleotide for all the clones. We calculate the background sequencing error rate by calculating the average fraction of non-reference alleles across all nucleotides where we did not attempt to introduce a mutation. Any site that has a significantly higher fraction of non-reference alleles (using a *P* value cutoff of 0.2 from a cumulative binomial test) is considered to have an unwanted mutation. A lenient *P* value cutoff (0.2 as opposed to the more traditionally used 0.05 or 0.01) implies more stringent filtering in this case because we want to eliminate type II errors i.e., we want to identify all unwanted mutations at the cost of discarding a few clones that actually do not have any unwanted mutations.

We identified an average of 4–5 unwanted point mutations per pool. The overall per-base point mutation rate of Phusion polymerase was calculated to be $\sim 10^{-4}$. NEB's advertised error rate for Phusion polymerase varies from 4.4–9.5×10^{-7} per PCR cycle. Since we perform 18 PCR cycles, the expected overall error rate is $\sim 10^{-5}$. Our calculated mutation is within an order of magnitude of this advertised error rate. It is slightly higher than the advertised rate as we use stringent filtering criteria as described above.

2.7.7 GFP Assay

All wild-type and mutant clones were moved into the pcDNA-DEST47 vector with a C-terminal GFP tag using automated Gateway LR reactions in a 96-well format. After

bacterial transformation, minipreps were prepared on a Tecan Freedom Evo 200, and DNA concentrations were determined by OD 260/280 with a Tecan Infinite M1000 plate reader in 96-well format. A 100 ng aliquot of each expression clone plasmid was used for transfection into HEK293T cells in 96-well plates using Lipofectamine 2000 (Invitrogen 11668019) according to the manufacturer's instructions. At approximately 48 hrs post-transfection, cells were processed with Tecan M1000. Fluorescence intensities were measured at 395 nm for excitation and 507 nm for emission, according to Invitrogen's manual. As negative controls, the fluorescence intensities corresponding to cells transfected with the empty vector were measured. The normalized fluorescence intensity was calculated as:

$$I_{norm} = I - I_{background}$$

where I corresponds to the measured intensity and $I_{background}$ corresponds to the average intensity of the empty vector controls for each plate. All I_{norm} values greater than K are considered to correspond to stable protein expression. K corresponds to the range (maximum – minimum) of background fluorescence intensities of the empty vector controls for each plate. For this study, all fluorescence intensity readings were also confirmed manually under a microscope. All transfection and GFP experiments were repeated 3 times.

2.7.8 Y2H Assay

Y2H was performed as previously described (Wang et al., 2012). All wild-type/mutant clones were transferred by Gateway LR reactions into our Y2H pDEST-AD and pDEST-DB vectors. All DB-X and AD-Y plasmids were transformed individually into

the Y2H strains *MAT*α Y8930 and *MAT***a** Y8800, respectively. Each of the DB-X *MAT*α transformants (wild-type and mutants) were then mated against corresponding AD-Y *MAT***a** transformants (wild-type and mutants) individually using automated 96-well procedures, including inoculation of AD-Y and DB-X yeast cultures, mating on YEPD media (incubated overnight at 30°C), and replica-plating onto selective Synthetic Complete media lacking leucine, tryptophan, and histidine, and supplemented with 1 mM of 3-amino-1,2,4-triazole (SC-Leu-Trp-His+3AT), SC-Leu-His+3AT plates containing 1 mg/l cycloheximide (SC-Leu-His+3AT+CHX), SC-Leu-Trp-Adenine (Ade) plates, and SC-Leu-Ade+CHX plates to test for CHX-sensitive expression of the *LYS2::GAL1-HIS3* and *GAL2-ADE2* reporter genes. The plates containing cycloheximide select for cells that do not have the AD plasmid due to plasmid shuffling. Growth on these control plates thus identifies spontaneous auto-activators (Walhout and Vidal, 2001). The plates were incubated overnight at 30°C and "replica-cleaned" the following day. Plates were then incubated for another three days, after which positive colonies were scored as those that grow on SC-Leu-Trp-His+3AT and/or on SC-Leu-Trp-Ade, but not on SC-Leu-His+3AT+CHX or on SC-Leu-Ade+CHX. Disruption of an interaction by a mutation was defined as at least 50% reduction of growth consistently across both reporter genes, when compared to Y2H phenotypes of the corresponding wild-type allele as benchmarked by 2-fold serial dilution experiments. All Y2H experiments were repeated 3 times.

2.7.9 Construction of plasmids

Wild-type *MLH1*, *HSPA8*, and *BRIP1* entry clones are from the human ORFeome v8.1 collection (Yang et al., 2011). Using Gateway LR reactions, wild-type *MLH1*, mutant *MLH1* (I107R), and GFP were transferred into the pMSCV-N-FLAG-HA-PURO vector (Behrends et al., 2010); *HSPA8* and *BRIP1* were transferred into the pcDNA-DEST40 vector that contains a C-terminal V5 tag (Invitrogen 12274-015).

2.7.10 Cell culture, co-immunoprecipitation, and Western blotting

HEK293T cells were maintained in complete DMEM medium supplemented with 10% FBS. Cells were transfected with Lipofectamine 2000 (Invitrogen) at a 6∶1 (μL/μg) ratio with DNA in 6-well plates and were harvested 24 hrs after transfection. Cells were gently washed three times in PBS and then resuspended using 200 μL 1% NP-40 lysis buffer [1% Nonidet P-40, 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1× EDTA-free Complete Protease Inhibitor tablet (Roche)] and kept on ice for 20 mins. Extracts were cleared by centrifugation for 10 mins at 13,000 rpm at 4°C. 15 μL EZview Red Anti-HA Affinity Gel (Sigma-Aldrich) and 100 μL protein lysate were used for each co-immunoprecipitation reaction. The samples were rotated gently at 4°C for 2 hrs. HA beads were then washed three times with protein lysis buffer, treated with 6× protein sample buffer, and subjected to SDS-PAGE. Proteins were then transferred from the gel onto PVDF (Amersham) membranes. Anti-HA (Sigma H9658), anti-V5 (Invitrogen 46-0705), anti-β-tubulin (Promega G7121), and anti-GFP (Santa Cruz sc-9996) antibodies were used at 1∶3,000 dilutions for immunoblotting analysis.

2.8 References

Arabidopsis Interactome Mapping Consortium (2011). Evidence for Network Evolution in an *Arabidopsis* Interactome Map. Science *333*, 601-607.

Araya, C.L., Fowler, D.M., Chen, W., Muniez, I., Kelly, J.W., and Fields, S. (2012). A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. Proceedings of the National Academy of Sciences *109*, 16858-16863.

Bandyopadhyay, A., Saxena, K., Kasturia, N., Dalal, V., Bhatt, N., Rajkumar, A., Maity, S., Sengupta, S., and Chakraborty, K. (2012). Chemical chaperones assist intracellular folding to buffer mutational variations. Nature Chemical Biology *8*, 238.

Behrends, C., Sowa, M.E., Gygi, S.P., and Harper, J.W. (2010). Network organization of the human autophagy system. Nature *466*, 68.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep Iii, P.W., and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nature Biotechnology *24*, 1429.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Research *28*, 235-242.

Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J.M., Murray, R.R., Roncari, L., de Smet, A.-S.*, et al.* (2009). An experimentally derived confidence score for binary protein-protein interactions. Nat Meth *6*, 91-97.

Cunningham, B., and Wells, J. (1989). High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. Science *244*, 1081-1085.

Das, J., Lee, H.R., Sagar, A., Fragoza, R., Liang, J., Wei, X., Wang, X., Mort, M., Stenson, P.D., Cooper, D.N.*, et al.* (2014). Elucidating Common Structural Features of Human Pathogenic Variations Using Large-Scale Atomic-Resolution Protein Networks. Human Mutation *35*, 585-593.

Das, J., Vo, T.V., Wei, X., Mellor, J.C., Tong, V., Degatano, A.G., Wang, X., Wang, L., Cordero, N.A., Kruer-Zerhusen, N.*, et al.* (2013). Cross-Species Protein Interactome Mapping Reveals Species-Specific Wiring of Stress Response Pathways. Sci Signal *6*, ra38-.

Finn, R.D., Marshall, M., and Bateman, A. (2005). iPfam: visualization of protein–protein interactions in PDB at domain and amino acid resolutions. Bioinformatics *21*, 410-412.

Fowler, D.M., Araya, C.L., Fleishman, S.J., Kellogg, E.H., Stephany, J.J., Baker, D., and Fields, S. (2010). High-resolution mapping of protein sequence-function relationships. Nature Methods *7*, 741.

Franzosa, E.A., and Xia, Y. (2011). Structural principles within the human-virus protein-protein interaction network. Proceedings of the National Academy of Sciences *108*, 10538-10543.

Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Altshuler, D., Shendure, J., Nickerson, D.A.*, et al.* (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216-220.

Hanks, S., Coleman, K., Reid, S., Plaja, A., Firth, H., FitzPatrick, D., Kidd, A., Méhes, K., Nash, R., Robin, N.*, et al.* (2004). Constitutional aneuploidy and cancer predisposition caused by biallelic mutations in BUB1B. Nature Genetics *36*, 1159.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences *106*, 9362-9367.

Houlston, R., Bevan, S., Williams, A., Young, J., Dunlop, M., Rozen, P., Eng, C., Markie, D., Woodford-Richens, K., Rodriguez-Bigas, M.A.*, et al.* (1998). Mutations in DPC4 (SMAD4) cause juvenile polyposis syndrome, but only account for a minority of cases. Human Molecular Genetics *7*, 1907-1912.

Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A.*, et al.* (2013). Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. Science *342*.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Machado, R.D. (2012). The Molecular Genetics and Cellular Mechanisms Underlying Pulmonary Arterial Hypertension. Scientifica *2012*, 17.

Massagué, J. (2008). TGFβ in Cancer. Cell *134*, 215-230.

Matsuura, S., Matsumoto, Y., Morishima, K.i., Izumi, H., Matsumoto, H., Ito, E., Tsutsui, K., Kobayashi, J., Tauchi, H., Kajiwara, Y.*, et al.* (2006). Monoallelic BUB1B mutations and defective mitotic-spindle checkpoint in seven families with premature chromatid separation (PCS) syndrome. American Journal of Medical Genetics Part A *140A*, 358-367.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M.*, et al.* (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research *20*, 1297-1303.

Nasim, M.T., Ogo, T., Ahmed, M., Randall, R., Chowdhury, H.M., Snape, K.M., Bradshaw, T.Y., Southgate, L., Lee, G.J., Jackson, I.*, et al.* (2011). Molecular genetic characterization of SMAD signaling molecules in pulmonary arterial hypertension. Human Mutation *32*, 1385-1389.

Pitt, J.N., and Ferré-D'Amaré, A.R. (2010). Rapid Construction of Empirical RNA Fitness Landscapes. Science *330*, 376-379.

Reece-Hoyes, J.S., Barutcu, A.R., McCord, R.P., Jeong, J.S., Jiang, L., MacWilliams, A., Yang, X., Salehi-Ashtiani, K., Hill, D.E., Blackshaw, S.*, et al.* (2011). Yeast one-hybrid assays for gene-centered human gene regulatory network mapping. Nature Methods *8*, 1050.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, Samuel J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R.*, et al.* (2014). A Proteome-Scale Map of the Human Interactome Network. Cell *159*, 1212-1226.

Roth, S., Sistonen, P., Salovaara, R., Hemminki, A., Loukola, A., Johansson, M., Avizienyte, E., Cleary, K.A., Lynch, P., Amos, C.I.*, et al.* (1999). SMAD genes in juvenile polyposis. Genes, Chromosomes and Cancer *26*, 54-61.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N.*, et al.* (2005). Towards a proteome-scale map of the human protein-protein interaction network. Nature *437*, 1173-1178.

Salehi-Ashtiani, K., Yang, X., Derti, A., Tian, W., Hao, T., Lin, C., Makowski, K., Shen, L., Murray, R.R., Szeto, D.*, et al.* (2008). Isoform discovery by targeted cloning, 'deep-well' pooling and parallel sequencing. Nature Methods *5*, 597.

Starita, L.M., Pruneda, J.N., Lo, R.S., Fowler, D.M., Kim, H.J., Hiatt, J.B., Shendure, J., Brzovic, P.S., Fields, S., and Klevit, R.E. (2013). Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. Proceedings of the National Academy of Sciences *110*, E1263-E1272.

Stein, A., Céol, A., and Aloy, P. (2011). 3did: identification and classification of domain-based interactions of known three-dimensional structure. Nucleic Acids Research *39*, D718-D723.

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

Stransky, N., Egloff, A.M., Tward, A.D., Kostic, A.D., Cibulskis, K., Sivachenko, A., Kryukov, G.V., Lawrence, M.S., Sougnez, C., McKenna, A.*, et al.* (2011). The Mutational Landscape of Head and Neck Squamous Cell Carcinoma. Science *333*, 1157-1160.

Suijkerbuijk, S.J.E., van Osch, M.H.J., Bos, F.L., Hanks, S., Rahman, N., and Kops, G.J.P.L. (2010). Molecular Causes for BUBR1 Dysfunction in the Human Cancer Predisposition Syndrome Mosaic Variegated Aneuploidy. Cancer Research *70*, 4891-4900.

Suzuki, Y., Kagawa, N., Fujino, T., Sumiya, T., Andoh, T., Ishikawa, K., Kimura, R., Kemmochi, K., Ohta, T., and Tanaka, S. (2005). A novel high-throughput (HTP) cloning strategy for site-directed designed chimeragenesis and mutation using the Gateway cloning system. Nucleic Acids Research *33*, e109.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

The Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330.

Vandenbroucke, I., Marck, H.V., Verhasselt, P., Thys, K., Mostmans, W., Dumont, S., Eygen, V.V., Coen, K., Tuefferd, M., and Aerssens, J. (2011). Minor variant detection in amplicons using 454 massive parallel pyrosequencing: experiences and considerations for successful applications. BioTechniques *51*, 67-77.

Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.-I.*, et al.* (2009). An empirical framework for binary interactome mapping. Nat Meth *6*, 83-90.

Vidal, M., Cusick, Michael E., and Barabási, A.-L. (2011). Interactome Networks and Human Disease. Cell *144*, 986-998.

Walhout, A.J.M., and Vidal, M. (2001). High-Throughput Yeast Two-Hybrid Assays for Large-Scale Protein Interaction Mapping. Methods *24*, 297-306.

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotech *30*, 159-164.

Yakhnin, A.V., Yakhnin, H., and Babitzke, P. (2012). Gel Mobility Shift Assays to Detect Protein–RNA Interactions. In Bacterial Regulatory RNA: Methods and Protocols, K.C. Keiler, ed. (Totowa, NJ: Humana Press), pp. 201-211.

Yang, X., Boehm, J.S., Yang, X., Salehi-Ashtiani, K., Hao, T., Shen, Y., Lubonja, R., Thomas, S.R., Alkan, O., Bhimdi, T.*, et al.* (2011). A public genome-scale lentiviral expression library of human ORFs. Nat Meth *8*, 659-661.

Yu, H., Braun, P., Yıldırım, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N.*, et al.* (2008). High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. Science *322*, 104-110.

Yu, H., Tardivo, L., Tam, S., Weiner, E., Gebreab, F., Fan, C., Svrzikapa, N., Hirozane-Kishikawa, T., Rietman, E., Yang, X.*, et al.* (2011). Next-generation sequencing to generate interactome datasets. Nat Meth *8*, 478-480.

Zhong, Q., Simonis, N., Li, Q.-R., Charloteaux, B., Heuze, F., Klitgord, N., Tam, S., Yu, H., Venkatesan, K., Mou, D.*, et al.* (2009). Edgetic perturbation models of human inherited disorders. Mol Syst Biol *5*.

CHAPTER 3

EXTENSIVE PROTEIN INTERACTION PERTURBATIONS BY HUMAN POPULATION VARIANTS ACROSS RARE AND COMMON ALLELE FREQUENCIES

3.1 Preface

Chapter 3 will be submitted for publication as "Fragoza, R.*, Das, J.*, Wierbowski, S.D., Liang, J., Liang, S., Beltran, J.F., Rivera-Erick, C.A., Ye, K, Wang, T.-Y., Mort, M., Stenson, P.D., Cooper, D.N., Keinan, A., Clark, A.G., Yu, H. Extensive protein interaction perturbations by human population variants across rare and common allele frequencies." where * indicates co-first author. The contents may be modified upon publication.

3.2 Abstract

Coding variants segregating in human populations are expected to be largely benign, with deleterious variation occurring principally at rare allele frequencies and limited to conserved genomic sites. The extent to which this deleterious variation burdens human genomes and the mechanisms by which these mutations exert their function, though, remains experimentally unexplored. To address this gap, we leveraged the ExAC database of 60,706 human exomes to investigate the functional impact 2,053 single nucleotide variants (SNVs) across 2,254 protein-protein interactions, functionally profiling 4,832 SNV-interaction pairs. Disruptive population variants occurred more prevalently at lower allele frequencies, were frequently depleted in essential genes,

and were strongly enriched at conserved genomic sites, underscoring the functional importance of the sites harboring disruptive variants. By incorporating these results with the site frequency spectrum determined from ExAC, we also find that 11.2% of missense variants per individual are expected to be disruptive. Moreover, we demonstrate how candidate disease-associated mutations can be identified through shared interaction perturbations between variants of interest and known disease mutations. In this manner, we expect our interactome perturbation study and its results to serve as an important template for providing contextual information for interpreting the molecular impact of coding variation on protein function.

3.3 Introduction

Recent explosive population growth has generated an excess of rare genetic variation segregating in human populations and very likely plays a role in the individual genetic burden of complex disease risk (Keinan and Clark, 2012). In agreement with this paradigm, large-scale whole-genome and whole-exome sequencing efforts have reported extensive genetic variation in human genomes segregating at very low and rare allele frequencies (Fu et al., 2013; Lek et al., 2016; Nelson et al., 2012; Tennessen et al., 2012; The 1000 Genomes Project Consortium, 2012; The UK10K Consortium, 2015). This excess of rare and low frequency coding SNVs is also predicted to impact protein function (Tennessen et al., 2012; The 1000 Genomes Project Consortium, 2012; The Genome of the Netherlands Consortium, 2014) in individual genomes; however, methods and metrics for estimating the functionality of coding SNVs vary widely, and there is no consensus measure on the number of functional variants per

genome (Henn et al., 2015). As such, a direct assessment of the functionality of coding SNVs in absence of a consensus metric could prove indispensable towards furthering our understanding of the impact that segregating genetic variation has on complex traits and human disease.

Biological processes are likely regulated through intricate networks of protein and macromolecular interactions, as opposed to single proteins acting by themselves (Vidal, 2001; Vidal et al., 2011). Hence, towards an improved understanding of the functional impact of human population variants on protein function, we capitalized on the ExAC dataset of coding variants for 60,706 human exomes (Lek et al., 2016) to systematically evaluate the impact 2,053 missense SNVs across 2,254 protein-protein interactions. We find that disruptive SNVs are significantly enriched at interaction interfaces, occur more prevalently and perturb more interaction partners at decreasing allele frequencies, and seldom result in unstable protein expression. Importantly, disruptive SNVs are strongly enriched at ancestrally conserved sites in the genome, underscoring the functional importance of the disruptive variants uncovered by our assays. Moreover, we also determined that on average 11.2% of coding SNVs per individual genome are expected to impact protein interactions, a rate much higher than indicated by previous reports (Tennessen et al., 2012; The 1000 Genomes Project Consortium, 2012; The Genome of the Netherlands Consortium, 2014). Unexpectedly, while we do observe an enrichment of functional SNVs at rare allele frequencies in agreement with previous literature (Nelson et al., 2012; Tennessen et al., 2012; The 1000 Genomes Project Consortium, 2012; The Genome of the Netherlands Consortium, 2014), we also find that 13.5% of common variants with minor allele

frequency (MAF) > 5.0% tested perturb protein interactions, signifying that many common variants are also functional (Gibson, 2011; Manolio et al., 2008).

We note that while the elevated disruption rates at common allele frequencies reported here are higher than previous reports (Tennessen et al., 2012; The 1000 Genomes Project Consortium, 2012; The Genome of the Netherlands Consortium, 2014), this result does not imply that common variation grossly impacts organism fitness. Rather, we emphasize that the context and genetic background in which this disruptive variation occurs is crucial towards properly interpreting the impact that such variation may have on organism-level phenotypes. Indeed, we find that disruptive variants are depleted in essential and haploinsufficient genes, target proteins with lower betweenness centrality, and reach higher MAFs principally on lowly conserved sites in the genome, strongly suggesting that the potentially deleterious impact of disruptive SNVs is frequently mitigated at the gene and network level. Nevertheless, cellular and organism-level phenotypes all stem from macromolecular perturbations. Moreover, the genetic background of an individual and its influence on complex traits and disease is determined by the cumulative impact of functional variation, including disruptive SNVs (Chow, 2015). Hence, a refined understanding of the molecular-level impacts of disruptive SNVs is imperative towards advancing our understanding of the impact segregating coding variants impose on human health and disease.

3.4 Results

3.4.1 Generating a resource of 2,053 single nucleotide variant clones

In an effort to investigate the impact of SNVs on protein-protein interactions, we

selected missense variants from three major databases: 1,712 variants from ExAC (Lek et al., 2016), 204 disease-associated mutations from HGMD (Stenson et al., 2014), and 169 somatic mutations in cancer form COSMIC (Forbes et al., 2011). Notably, over half of variants found in ExAC are singletons. Consequently, to avoid oversampling rare population variants, we randomly selected alleles in ExAC across defined MAF bins ranging from singletons to very common alleles (**Figure 3.4-1**). We further required that corresponding genes for selected alleles were (1) present in the hORFeome (The MGC Project Team, 2009; Yang et al., 2011) and (2) were amenable to our version of the Y2H assay, which was used to generate the most comprehensive binary human interactome available (Rolland et al., 2014; Rual et al., 2005; Venkatesan et al., 2009; Yu et al., 2011). We then generated single clones for all selected variants using Clone-seq, a massively parallel pipeline for large-scale, site-directed mutagenesis (Wei et al., 2014). In this manner, we produced a resource of sequence-validated single-colony clones for 2,053 SNVs spanning 891 wild-type genes.

3.4.2 Disruptive coding SNVs occur extensively across wide allele frequency ranges in human genomes

Alterations to protein-protein interactions can have deleterious consequences to organism fitness (Schoenrock et al., 2017). Indeed, disease-associated mutations frequently function through perturbations to specific protein-protein interactions (Sahni et al., 2015; Wang et al., 2012; Zhong et al., 2009) and hence coding variation at interaction interfaces is predominantly rare (Khurana et al., 2013) and subject to

evolutionary constraint (Guharoy and Chakrabarti, 2005; Mintseris and Weng, 2005).

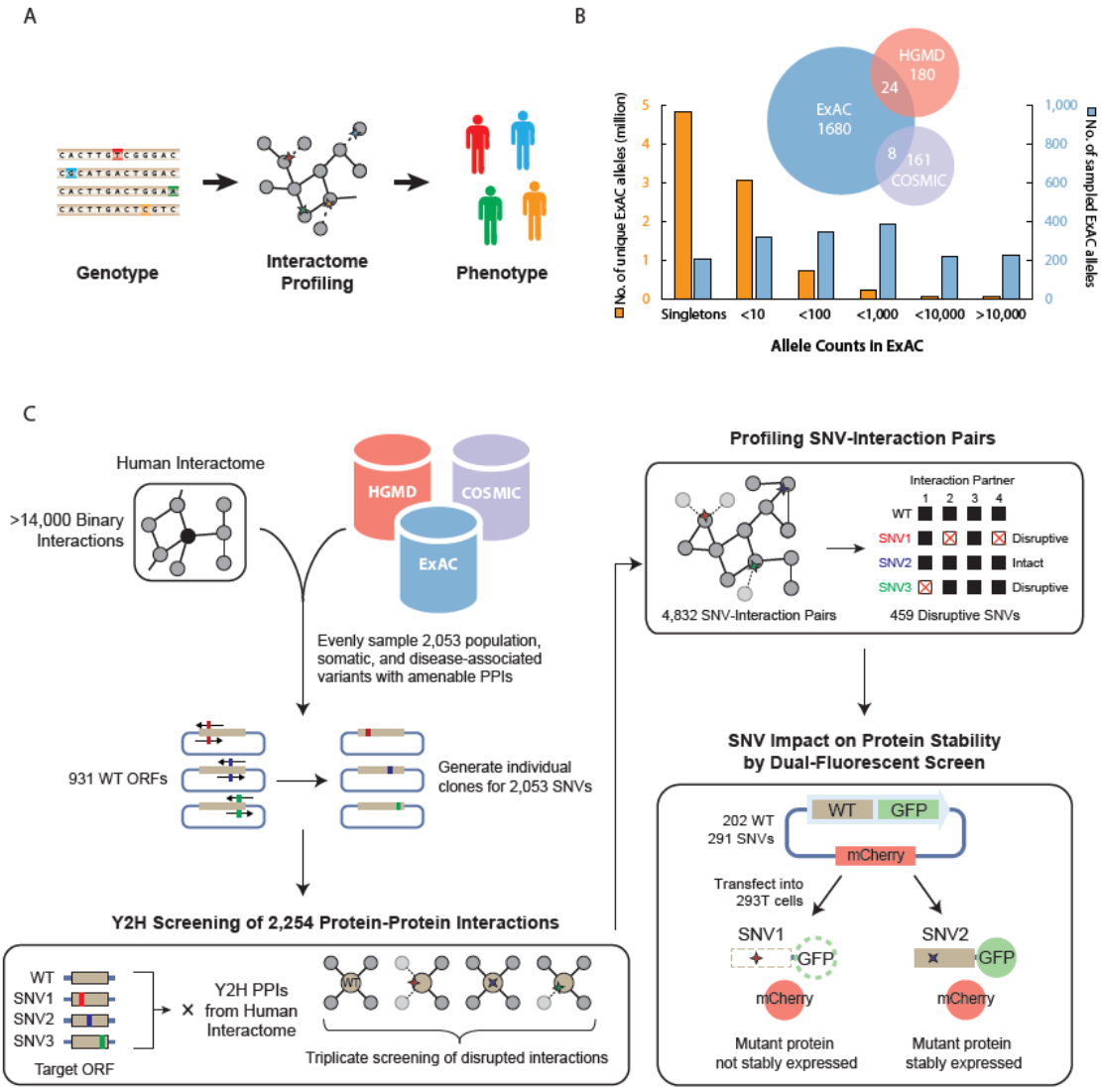


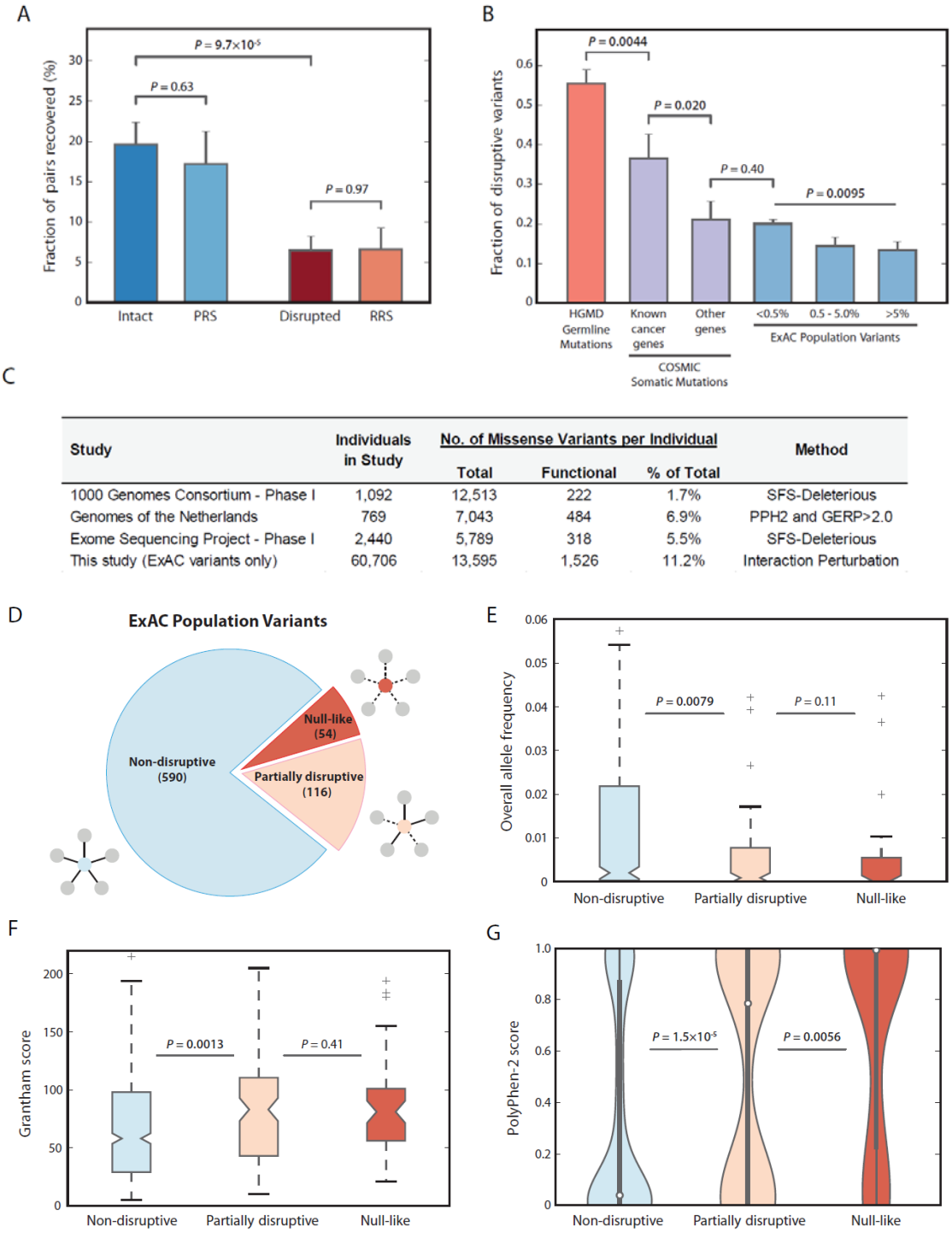
Figure 3.4-1 A pipeline for surveying the impact of 2,053 SNVs on protein-protein interactions. *(A) Phenotypic consequences of coding variants in human genotypes can be interpreted as products of protein-protein interaction perturbations in the interactome. (B) Over half of all unique missense variants in ExAC are singletons. To avoid oversampling very rare variants from ExAC, 1,712 ExAC variant were selected across a wide range of allele frequencies. 204 disease-associated mutations listed in HGMD and 169 cancer somatic mutations from COSMIC were also examined. (C) Pipeline for testing functional impact of 2,053 SNVs on protein interactions and stability impact of 291 population variants by dual-fluorescence screen.*

In contrast, common variation is expected to be largely neutral and therefore unlikely to harbor extensive functional variation (Adzhubei et al., 2010; Gorlov et al., 2008; Maher et al., 2012). Nonetheless, there are notable exceptions, including APOE-epsilon 4 (Corder et al., 1993; Deary et al., 2002; Strittmatter et al., 1993) and P12A polymorphism of PPARG (Florez et al., 2007; Robitaille et al., 2003). Indeed, the extent to which MAF indicates whether an allele is disruptive to protein interactions remains systematically unexplored. Hence, to systematically identify functionally relevant SNVs across rare to common allele frequencies, we performed yeast two-hybrid (Y2H) experiments for 2,053 missense SNVs tested across 2,254 corresponding protein interactions. In total, we identified 459 interaction-disrupting SNVs, including 309 disruptive ExAC variants, tested across 4,832 SNV-interaction pairs.

To validate the quality of our SNV-interaction network, we performed an orthogonal Protein Complementation Assay (PCA) (Das et al., 2014a) in human 293T cells to retest a representative subset of 401 disrupted and non-disrupted SNV-interactions pairs from ExAC. SNV-disrupted interactions retested at a rate highly similar to a negative reference set of 92 randomly selected ORF pairs while non-disrupted interactions retested at a rate statistically indistinguishable from a positive reference set of 92 well-established protein interactions (Braun et al., 2009; Venkatesan et al., 2009) (**Figure 3.4-2A**, **Supplementary Figure C.1-2A**). Taken together, our PCA retest demonstrates the reproducibility of our dataset and reinforces the high quality of our SNV-interaction network.

Next, we partitioned all tested population variants into three overall allele

frequency bins spanning from rare (MAF ≤ 0.5%), intermediate (0.5% < MAF ≤ 5.0%) to common ExAC alleles (MAF > 5%) and then calculated the fraction of variants that disrupted one or more protein interactions per MAF bin. We find that the fraction of disruptive variants decreased inversely with respect to allele frequency (P = 0.0095 by chi-square test), which agrees with expectations (Keinan and Clark, 2012; Tennessen et al., 2012; The 1000 Genomes Project Consortium, 2012); however, this decrease was not precipitous. 13.5% of common variants with MAF > 5% were still disruptive while 20.0% of rare variants with MAF ≤ 0.5% were measured as disruptive, a less than twofold increase in disruption rate despite at least an order of magnitude difference in allele frequency (**Figure 3.4-2B**). Considering the pervasiveness of common variants across populations, this result implies that disruptive coding variants are segregating across numerous individuals. As such, we then applied these binned-MAF disruption rates and the site frequency spectrum for missense variants calculated from ExAC to determine the fraction of missense variants per individual genome (**Appendix C.2.1**; Materials and Methods). We found that out of an average of 13,595 missense variants per genome, 1,526 variants (11.2%) are expected to disrupt protein interactions, a figure notably higher than indicated by previous estimates (**Figure 3.4-2C, Supplementary Tables C.3-1, C.3-2, and C.3-3**), though the extent to which interaction disruptions result in cellular phenotypes, particularly for common variants, remains undetermined.



Study	Individuals in Study	No. of Missense Variants per Individual			Method
		Total	Functional	% of Total	
1000 Genomes Consortium - Phase I	1,092	12,513	222	1.7%	SFS-Deleterious
Genomes of the Netherlands	769	7,043	484	6.9%	PPH2 and GERP>2.0
Exome Sequencing Project - Phase I	2,440	5,789	318	5.5%	SFS-Deleterious
This study (ExAC variants only)	60,706	13,595	1,526	11.2%	Interaction Perturbation

Figure 3.4-2 The frequency of observing a disruptive allele is inversely proportional to allele frequency. *(A) Fraction of protein pairs recovered by PCA for disrupted and intact interactions in comparison to positive and random reference sets (PRS and RRS). P values by two-tailed Z-test. (B) Fraction of disruptive variants in ExAC (blue) across three overall allele*

frequency ranges (i) < 0.5%, (ii) 0.5 - 5.0%, and (iii) > 5.0%. P value by chi-square test. Fraction of disruptive somatic mutations in COSMIC (purple) in known cancer-affiliated genes or other genes and fraction of disruptive germline disease-associated genes from HGMD (red) are also shown. P values by one-tailed Z-test. (C) Reported number of functional missense variants per individual genome varies extensively across different studies. (D) ExAC variants tested against ≥ 2 interactions further partitioned into three disruption categories. Distribution of (E) overall allele frequency, (F) Grantham scores, and (G) PolyPhen-2 scores across three disruption categories. Error bars in (A) and (B) indicate +SE of proportion. Thick black bars in (G) are the interquartile range, white dots display the median, and extended thin black lines represent 95% confidence intervals. P values in (E), (F), and (G) by one-tailed U-test. Significant P values in ***bold****.*

Regardless, our results indicate that many variants show some measure of functionality at least at the molecular level.

To add further context to our disruption rate analysis, we also examined the fraction of cancer somatic mutations that disrupt interactions and found that 36.5% of somatic mutations located in genes with established roles in cancer progression were disruptive (**Figure 3.4-2B**; Materials and Methods). Notably, this fraction decreased significantly to 21.2% for somatic mutations located in all other genes ($P = 0.020$ by one-tailed *Z*-test), a figure that well approximates the 20.0% disruption rate observed for rare (MAF ≤ 0.5%) ExAC alleles but is significantly reduced in comparison to somatic mutations in known cancer genes. In stark contrast, 55.4% of tested disease-associated mutations were measured as disruptive (**Figure 3.4-2B**).

Collectively, these trends in disruption rate suggest that somatic mutations in genes with no known associations to cancer oncogenesis behave effectively as rare variants. Genetic context in which a mutation occurs is therefore crucial. Consequently, disruption rates for somatic mutations skew closer towards disease-

associated mutations if they occur in known cancer genes.

Interaction-perturbing population variants are not equally disruptive. If we focus only on proteins with multiple interaction partners, the disruptiveness of a variant can be categorized by measuring the fraction of corresponding protein interactions disrupted by a particular allele. Accordingly, ExAC variants that leave all interactions intact were categorized as non-disruptive, variants that disrupt a subset of interaction partners were categorized as partially disruptive, and variants that disrupt all tested protein interactions were categorized as null-like (**Figure 3.4-2D**). Across these three categories, overall allele frequency for tested variants in ExAC decreased significantly from 0.20% for non-disruptive to 0.086% for partially-disruptive variants ($P = 0.0079$ by one-tailed *U*-test) then nominally to 0.024% for null-like alleles (**Figure 3.4-2E**). Markedly, though, null-like variants were significantly depleted at common allele frequencies (MAF > 5.0%) relative to intermediate and rare allele frequencies (MAF ≤ 5.0%), occurring at a 3.9% and 7.9% rate, respectively ($P = 0.045$ by one-tailed *Z*-test; **Supplementary Figure C.1-2A**). As such, this twofold depletion at common allele frequencies may indicate that null-like variants are more likely to be deleterious and may undergo stronger negative selection than partially disruptive variants as a result. Segregating between both classes of disruptive variants through interaction perturbation assays may therefore provide valuable insight towards identifying phenotypically deleterious mutations.

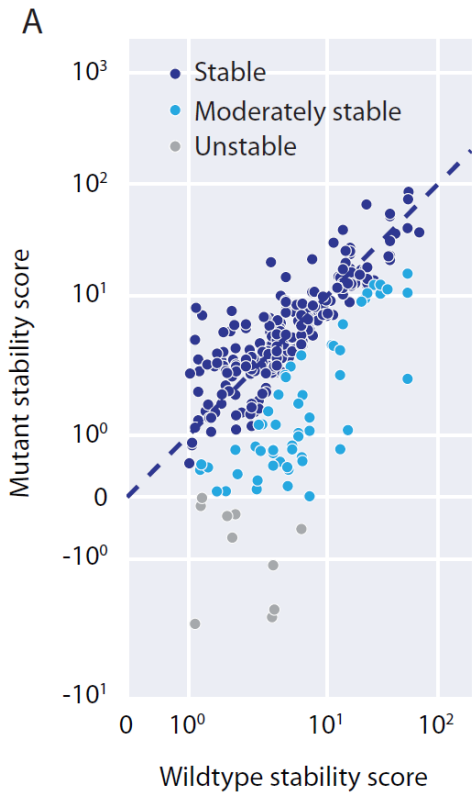
Highly dissimilar amino acid substitutions at conserved interaction sites should perturb protein interactions more frequently than relatively similar amino acid substitutions (Lockless and Ranganathan, 1999). Consequently, we expect that

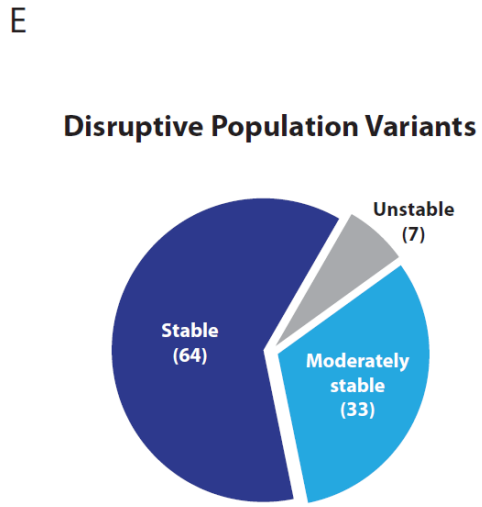
Grantham score, a biochemical measure quantifying the dissimilarity between amino acid residues (Grantham, 1974), for non-disruptive variants to be lower than for disruptive variants, particularly so for null-like variants which disrupt all tested protein interactions. While we found that Grantham score for partially disruptive variants was significantly higher than for non-disruptive variants (Medians = 83 and 58, respectively; $P = 0.0013$ by one-tailed *U*-test), no significant difference in Grantham score was observed between partially disruptive and null-like variants (**Figure 3.4-2F**). This result indicates that the contrasting interaction perturbation phenotypes observed between partially disruptive and null-like variants is not a consequence of more radical amino acid substitutions and that other selection or network-based properties may be driving these contrasting phenotypes instead. Indeed, conservation-based functional prediction algorithms including PolyPhen-2 (Adzhubei et al., 2010) and MutPred2 (Pejaver et al., 2017) show significant increases in the likelihood that a variant is deleterious across non-disruptive, partially disruptive, and null-like disruption categories (**Figure 3.4-2G**, **Supplementary Figure C.1-2B**). Collectively, these results imply that population variants that disrupt multiple interaction partners are, as a whole, under stronger functional constraint and occur at rarer allele frequencies as a result. However, despite this elevated constraint on null-like variants, understanding the context in which this constraint is imposed is imperative. For instance, if a null-like variant mutates an interface residue of a hub protein with only a single interaction interface, the corresponding site may still be conserved, despite occurring on a protein not likely to be essential (Kim et al., 2006). This topic is explored further in later sections.

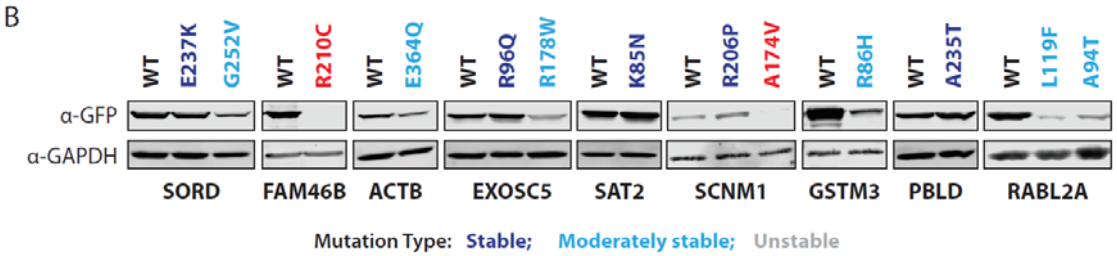
3.4.3 Missense variants seldom result in unstable protein expression

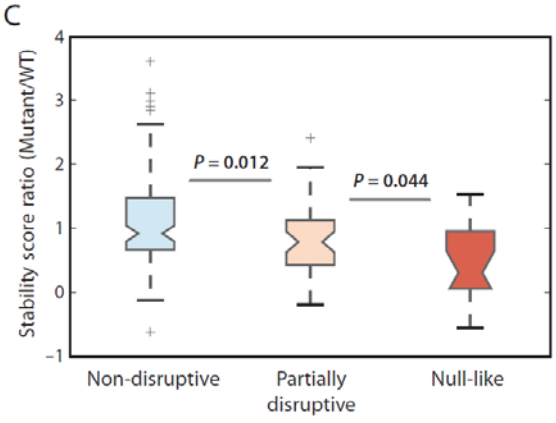
Mutations can disrupt interactions through local perturbations to specific interaction interfaces or by destabilizing protein folding as a whole (Zhong et al., 2009). To distinguish between these two mechanisms of disruption, we developed a dual fluorescence-based screening assay to survey the impact of interaction-disruptive variants on protein folding. To set-up our dual fluorescent screen, a subset of wild-type ORFs that are stably expressed when tagged with GFP and their corresponding ExAC variants were cloned into pDEST-DUAL, a GFP-tagged expression vector that co-expresses untagged mCherry (Methods). Wild-type and mutant ORFs were then transfected into 293T cells to test for mutation-induced changes in protein expression in human cells on 96-well plate scales (**Fig 3.4-1C**). GFP expression levels for transfected wild-type and mutant samples were then measured and normalized with respect to mCherry expression levels, reported as wild-type and mutant stability scores (**Figure 3.4-3A**), to determine the impact of 291 ExAC variants on protein folding.

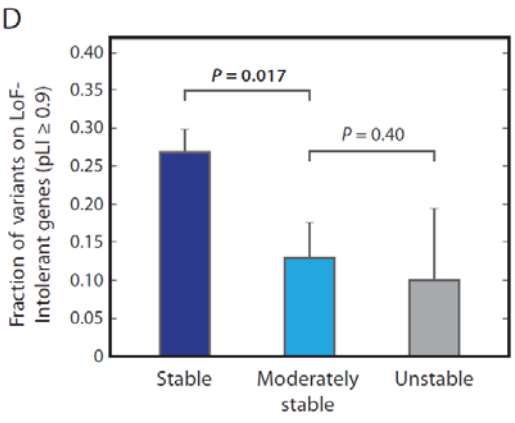
We next grouped mutant proteins into three categories of expression. Specifically, if the ratio between mutant and wild-type stability score falls below 0.5, indicating that the mutant protein is still expressed but at markedly reduced levels, we categorize the mutant protein as moderately stable. If, though, mutant protein expression drops below plate reader-detectable measures, as indicated by a mutant stability score < 0, we instead categorize the mutant protein as unstable.


A
Stable
Moderately stable
Unstable
Mutant stability score
Wildtype stability score


E
Disruptive Population Variants
Unstable (7)
Stable (64)
Moderately stable (33)


B
WT E237K G252V
WT R210C
WT E364Q
WT R96Q R178W
WT K85N
WT R206P A174V
WT R86H
WT A235T
WT L119F A94T
α-GFP
α-GAPDH
SORD
FAM46B
ACTB
EXOSC5
SAT2
SCNM1
GSTM3
PBLD
RABL2A
Mutation Type: Stable; Moderately stable; Unstable


C
Stability score ratio (Mutant/WT)
P = 0.012
P = 0.044
Non-disruptive
Partially disruptive
Null-like


D
Fraction of variants on LoF-Intolerant genes (pLI ≥ 0.9)
P = 0.017
P = 0.40
Stable
Moderately stable
Unstable

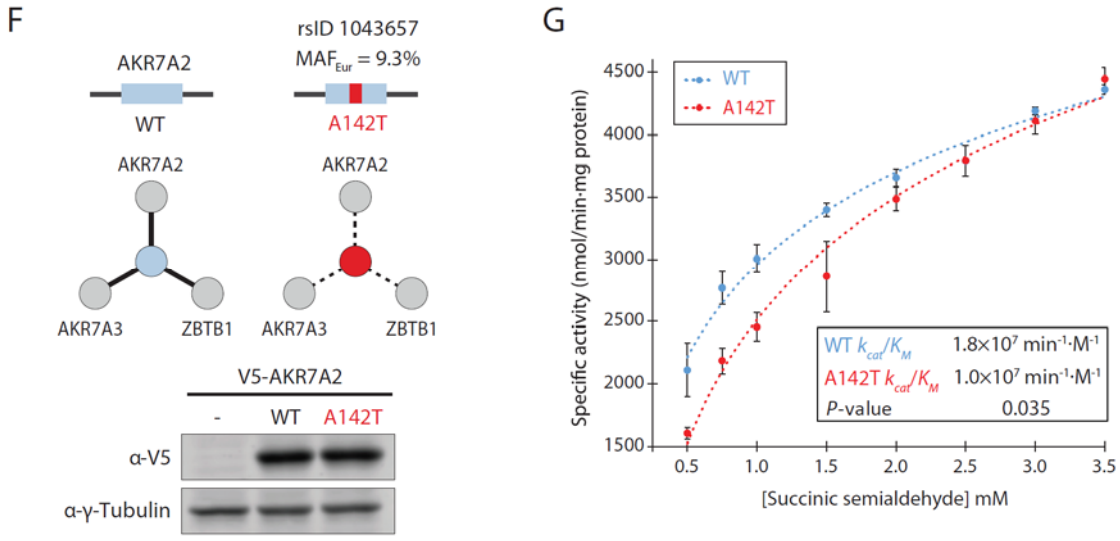


Figure 3.4-3 Disruptive population variants seldom result in unstable protein expression. *(A) DUAL-FLOU protein stability scores for 291 wild-type:variant pairs. (B) Western blots for representative wild-type:variant pairs across three stability categories detected using α-GFP. α-GAPDH was used as a loading control. (C) Ratio of mutant to wild-type stability score corresponding to non-disruptive (n = 108), partially disruptive (n = 46), and null-like variants (n = 17). (D) Fraction of variants residing on essential genes (pLI ≥ 0.9) for stable (n = 209), moderately stable (n = 54), and unstable (n = 10) protein stability categories. (E) Distribution of interaction-disruptive ExAC variants across three stability categories. (F) Diagram of interactions disrupted by null-like AKR7A2_A142T variant. Stable expression of V5-tagged AKR7A2 was validated by Western blot using α-V5. α-γ-Tubulin was used as a loading control. (G) In vitro specific activities of purified recombinant AKR7A2 wild-type and A142T using succinic semialdehyde substrate. Fitted curves (dashed lines) are shown for wild-type and A142T. P value by one-tailed t-test. Error bars indicate ±SE of mean at eight different substrate concentrations. Error bars in (C) and (D) indicate +SE of proportion. P values in (C) and (D) by one-tailed U-test. Significant P-values in* ***bold****.*

Mutant proteins above both thresholds are scored as stable otherwise (**Appendix C.2.1**). We note that our stable, moderately stable, and unstable demarcations correspond well with western blot intensity (**Figure 3.4-3B**).

Loss of protein expression is a strong molecular phenotype. Severe cell- and organism-level phenotypes can manifest from loss of protein expression, particularly if

the protein lost is essential. Hence, we used pLI (Lek et al., 2016) – a measure for how intolerant a gene is to loss-of-function (LoF-Intolerant) mutations – to investigate how often destabilizing protein variants occurred on essential genes across our three stability categories. Interestingly, we found that the fraction of variants on LoF-Intolerant genes (pLI ≥ 0.9) decreased significantly from 27% to 13% for stable and moderately stable mutant proteins, respectively ($P = 0.017$ by one-tailed *U*-test; **Figure 3.4-3D**), while only one unstable singleton mutation, was found. This depletion for moderately stable and unstable mutations on essential genes strongly implies that destabilizing mutations largely persist in genomic regions with little influence on organism fitness.

A destabilizing mutation in any other region would be more susceptible to negative selection and expected to occur at lower allele frequencies as result. In agreement, we observe that the overall allele frequency decreases from 4.2% for stable protein variants to 0.82% for moderately stable variants ($P = 0.033$ by one-tailed *U*-test; **Supplementary Figure C.1-3A**) and nominally to 0.17% for unstable.

To determine if protein stability corresponds with interaction-disruptive phenotypes, we compared the distribution of stability scores across tested alleles from non-disruptive, partially disruptive, and null-like categories. We found that the ratio of mutant to wild-type stability scores lowers significantly between non-disruptive and partially disruptive variants ($P = 0.012$ by one-tailed *U*-test) and between partially disruptive to null-like variants ($P = 0.044$ by one-tailed *U*-test; **Figure 3.4-3C**); however, despite this downward trend in stability score across interaction-disruptive categories, relatively few unstable variants were detected. Indeed, we found only

seven cases in which an interaction-disruptive variant resulted in unstable mutant protein expression (**Figure 3.4-3E**). As such, we conclude that disruptive population variants may be better explained by local structural perturbations that disrupt specific protein interactions as opposed to destabilizing protein stability as a whole.

Although protein-destabilizing mutations are depleted among essential genes, likely mitigating the impact an otherwise damaging variant would have on organism fitness, this depletion does not imply that variants on stably expressed non-essential genes are without functional impact. Nor must a population variant occur at rare allele frequency to be functionally consequential. To demonstrate this, we examined a null-like amino acid mutation, A142T, on a non-essential gene, AKR7A2 (pLI = 0), that occurs at common allele frequencies, particularly in African populations in ExAC (MAF_{Afr} = 9.3%) and does not impact protein expression (**Figure 3.4-3F**). AKR7A2 is an NADPH-dependent aldo1-keto reductase that catalyzes the reduction of succinic semialdehyde (SSA) to gamma-hydroxy butyrate (GHB), an important activity in the degradation pathway for the inhibitory neurotransmitter GABA (Lyon et al., 2007) (**Supplementary Figure C.1-3B**).

To explore the potential enzymatic impact of AKR7A2_A142T, we purified recombinant wild-type and mutant AKR7A2 protein to test for changes in NADPH-dependent turnover of SSA (Methods). Accordingly, we found that k_{cat}/K_M decreased from $1.8\times10^7\ min^{-1} \cdot M^{-1}$ for wild-type protein to $1.0\times10^7\ min^{-1} \cdot M^{-1}$ for AKR7A2_A142T (P = 0.035 by one-tailed t-test, **Figure 3.4-3G**). In addition to impacting SSA turnover, there is further *in vitro* evidence that the A142T variant of AKR7A2 impacts drug metabolism (Bains et al., 2010). While the high allele

frequency of this variant strongly implies that A142T has a minimal fitness impact in individuals, disease-associated mutations in the GABA degradation pathway to which AKR7A2 belongs indicate that AKR7A2_A142T could still be deleterious in certain genetic backgrounds. For example, missense mutations that impair ABAT and SSADH activity, enzymes immediately upstream of AKR7A2, can result in severe human neurological disorders (Akaboshi et al., 2003; Medina-Kauwe et al., 1998; Tsuji et al., 2010). Hence, we postulate that AKR7A2_A142T may indeed be deleterious in genetic backgrounds with lowered ABAT or SSADH activity, reinforcing that genetic background is a crucial component in determining whether disruptive variants can truly impact organism fitness. Similarly, our interactome perturbation study and its results will provide contextual information for many disease mutations, especially those with partial penetrance.

3.4.4 Disruptive population variants are enriched on conserved protein sites

Deleterious variants are more likely to be population-specific in comparison to neutral mutations since selection limits their capacity to spread across populations over time (Marth et al., 2011). As such, we explored how the fraction of disruptive variants changes across different populations sampled by ExAC. We found that disruptive variants are significantly less likely to be found across all ExAC populations in comparison to non-disruptive variants ($P = 0.014$ by one-tailed Z-test) and instead are enriched within a limited number of populations (1-3 populations: $P = 0.036$ by one-tailed Z-test, **Figure 3.4-4A**). We note, though, that the fraction of variants that are disruptive does not vary significantly across any individual population

(**Supplementary Figure C.1-4A**). Negative selection takes time to remove disruptive variants from a population. As such disruptive variants that are older in age would be removed from a population while disruptive variants that are younger would persist likely as private variants or as variants limited to a few populations. Our results therefore indicate that disruptive variants may be a consequence of younger variants recently emerged in human genomes.

The functional importance of a particular genomic site or protein residue can also be strongly inferred through sequence conservation with closely related ancestral DNA (Boffelli et al., 2003) or across multiple species (Asthana et al., 2007; Cooper et al., 2003; Margulies et al., 2003). Indeed, variants at sites in which the minor allele is ancestral are less likely to undergo purifying selection (Zhu et al., 2011), which may indicate that such sites may be less functionally relevant. In agreement, we observed a near two-fold increase in the fraction of disruptive variants at sites in which the ancestral allele matches the major allele as opposed to matching the tested minor allele (18.6% to 9.6%, respectively; $P = 0.025$ by one-tailed Z-test, **Figure 3.4-4B**). Moreover, while sites in which the major allele is ancestral are more conserved than sites in which the minor allele is ancestral ($P = 0.016$ by one-tailed U-test, **Supplementary Figure C.1-4B**), this difference intensified dramatically when we directly compared protein residues corresponding to disruptive alleles to those for non-disruptive alleles ($P = 3.0\times10^{-10}$ by one-tailed U-test, **Figure 3.4-4C**). Sequence conservation across multiple species is a foundational metric for identifying functional variants (Asthana et al., 2007; Cooper et al., 2003; Margulies et al., 2003), and hence, this result strongly implies that disruptive variants are targeting evolutionarily

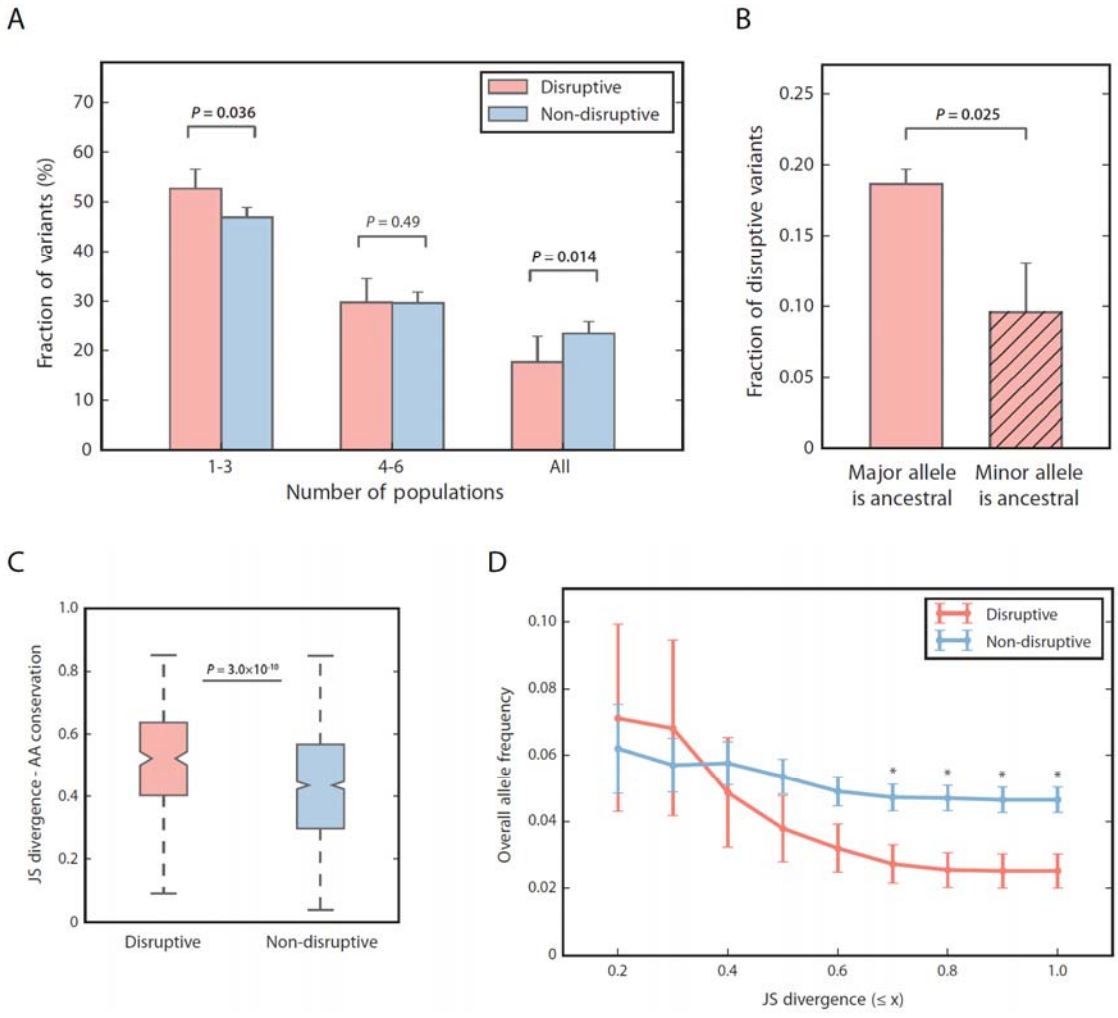
important genomic sites.

Deleterious mutations that occur on conserved protein residues are likely to undergo purifying selection while mutations to non-conserved protein residues can persist more readily in absence of selective constraint (McDonald and Kreitman, 1991). To search for evidence of purifying selection on disruptive alleles, we binned both disruptive and non-disruptive alleles by their corresponding Jensen-Shannon divergence (JSD) scores, an amino acid-based metric for conservation, and then compared the mean overall allele frequency of variants per JSD scoring bin (Methods). We found that while allele frequency for both disruptive and non-disruptive variants was comparable at low JSD conservation scores, mean allele frequency for disruptive variants strongly decreased across increasing JSD cutoffs. In contrast, non-disruptive variants decreased only mildly (**Figure 3.4-4D**). A similar pattern was also observed using a genomic, as opposed to amino acid-based, measure for conservation, phyloP (Pollard et al., 2010) (**Supplementary Figure C.1-4C**).

These corroborating results may suggest that allele frequency for variants at lowly conserved sites is principally governed by random genetic drift, regardless of whether the allele is disruptive or not. In contrast, negative selection acts more readily on disruptive alleles that occur at highly conserved sites, where deleterious mutations can strongly reduce organism fitness. These results further suggest that prioritizing disruptive rare variants at conserved sites, as opposed to rare variants at any sites, may yield higher resolution in identifying trait- or disease-associated mutations.

Although enriched at conserved coding sites, disruptive variants are not strictly constrained to such regions. Therefore we investigated whether disruptive variants are

also enriched in genomic regions undergoing positive selection. We applied Fay and Wu's H to measure positive selection genome-wide using Phase 3 data from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015) then examined where disruptive variants reside with respect to genomic regions undergoing significant positive selection (Methods).



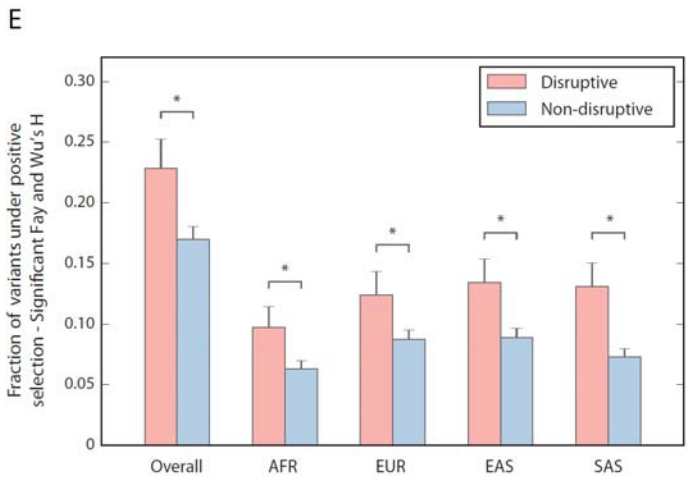


Figure 3.4-4 Disruptive alleles occur predominately at conserved genomic sites. *(A) Fraction of variants that occur in up to three populations, between 4-6 populations, or all seven populations in ExAC, partitioned into disruptive and non-disruptive categories. (B) Fraction of disruptive variants found at sites in which the major allele matches the ancestral chimp allele (purple, n = 1,464) and at sites in which the tested minor allele matches the ancestral chimp allele (yellow, n = 73). (C) Distribution of Jensen-Shannon Divergence scores for amino acid residues at sites corresponding to disruptive and non-disruptive alleles. Larger scores indicate a more conserved sites. (D) Relationship between conservation and overall allele frequency for disruptive and non-disruptive variants examined across increasing cutoff scores for JS divergence scores. Error bars indicate ±SE of mean. (E) Fraction of disruptive variants in genomic regions under positive selection indicated by a significant value for Fay and Wu's H shown across four different population groups and across overall population. Error bars in (A), (B), and (E) indicate +SE of proportion. P values in (A), (B), and (E) by one-tailed Z-test. P values in (C) and (D) by one-tailed U-test. Significant P values in* ***bold****. * $P < 0.05$.*

Unexpectedly, we found that disruptive variants across 1000 Genomes population groups were enriched within regions undergoing significant positive selection across the overall population as well as within specific population subgroups (**Figure 3.4-4E**). Genomic regions under positive selection have played crucial roles in the evolution of human phenotypic traits (Vallender and Lahn, 2004). Enrichment

for disruptive variants within such regions may therefore be a reflection of the functional importance of genomic regions under positive selection. However, not all genomic regions under positive selection can be readily ascribed to known molecular functions. Under our interaction perturbation framework, we can potentially better understand how positive selection emerges at a molecular level.

3.4.5 Disruptive variants are depleted among genes that strongly impact organism fitness

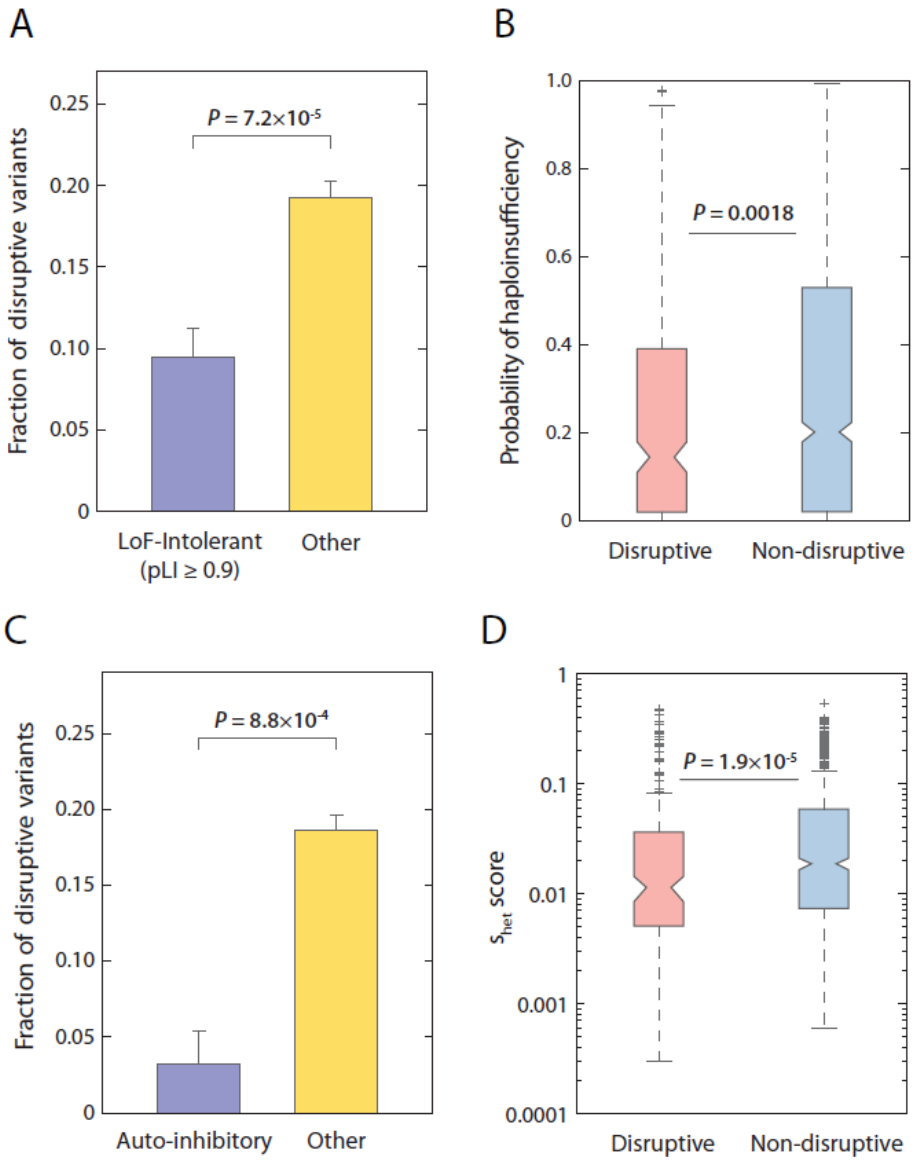
Our data revealed that disruptive protein variants occur predominately on conserved residues (**Figure 3.4-4C**), reflecting the functional importance of disrupted protein sites and that the overall allele frequency of disruptive variants decreases across increasingly conserved sites (**Figure 3.4-4D**). This decreased occurrence at highly conserved sites implies that disruptive alleles occur less frequently in genes crucial for proper cell function and instead persist in coding regions less susceptible to purifying selection. In agreement, we found that disruptive alleles are strongly depleted within LoF-Intolerant genes (pLI $\geq$ 0.9) in comparison to other genes (9.5% and 19.2%, respectively; $P = 7.2\times10^{-5}$ by one-tailed *Z*-test, **Figure 3.4-5A**). Likewise, we also found that genes harboring disruptive alleles are significantly less likely to be haploinsufficient (Bartha et al., 2017; Bartha et al., 2015) in comparison to those with non-disruptive alleles (Medians = 0.14 and 0.20, respectively; $P = 0.0018$ by one-tailed *U*-test, **Figure 3.4-5B**). Lastly, the fraction of disruptive variants that occur on disease-associated genes known to be autosomal dominant is significantly reduced in comparison to those on other genes (Online Mendelian Inheritance of Man OMIM;

Singh et al., 2014) (10.9% and 18.5%, respectively; $P = 0.033$ by one-tailed Z-test, **Supplementary Figure C.1-5A**). All three gene categories require only a single loss-of-function mutation to exert a deleterious cellular effect and hence selection acts more readily on disruptive alleles that occur in these often essential genes.

While essential genes are can be characterized by their intolerance to loss-of-function mutations (Bartha et al., 2017; Lek et al., 2016; Wang et al., 2014), some gene classes, particularly oncogenes and proteins with autoinhibitory protein folds such as SNARE and WASP proteins (Pufall and Graves, 2002), can instead be classified by their susceptibility to deleterious gain-of-function mutations (Pufall and Graves, 2002; Singh et al.). Such mutations are perilous and should occur infrequently because they often result in constitutively active mutants with dominant deleterious phenotypes (Pufall and Graves, 2002; Singh et al.). Accordingly, we found that the fraction of disruptive alleles is significantly depleted within known oncogenes (Bozic et al., 2010; Singh et al., 2014) in comparison to other genes (14.5% and 19.3%, respectively; $P = 0.011$ by one-tailed Z-test, **Supplementary Figure C.1-5B**) and particularly so in gene-encoded proteins with autoinhibitory folds (Singh et al.; Singh et al., 2014) in comparison to other genes (3.2% and 18.6%, respectively; $P = 8.8\times10^{-4}$ by one-tailed Z-test, **Figure 3.4-5C**). Despite a higher propensity for acquiring deleterious mutations, selection acts readily on deleterious mutations in such genes and hence disruptive mutations are strongly depleted within both gene categories.

Disruptive mutations are depleted in essential, haploinsufficient genes and genes susceptible to deleterious gain-of-function mutations since selection can act readily to deplete deleterious alleles in both gene categories. Disruptive mutations may

therefore persist in genes less impactful to cell viability where selection acts less readily. Indeed we found that genes harboring disruptive variants corresponded with significantly lower selection coefficients than genes harboring non-disruptive variants (Medians = 0.011 and 0.019, respectively; $P = 1.9\times10^{-5}$ by one-tailed *U*-test, **Figure 3.4-5D**). This result suggests that abundant interaction-disruptive alleles in human populations persist as a result of weaker purifying selection acting in non-essential regions of human genomes.



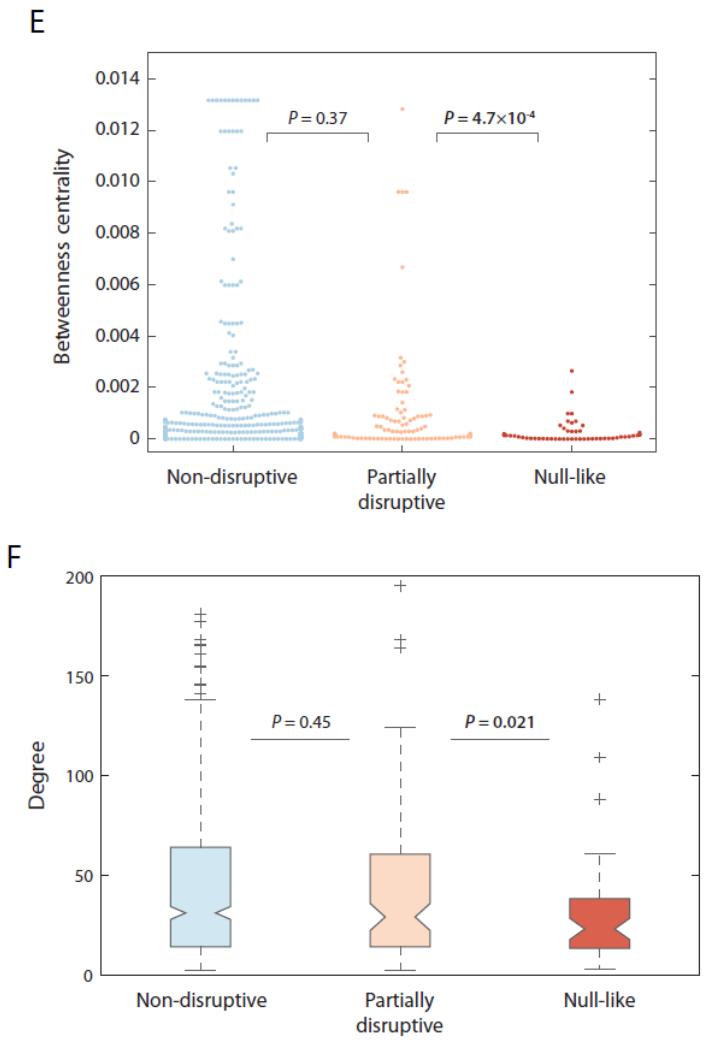


Figure 3.4-5 Disruptive variants are depleted among essential genes and genes prone to deleterious mutations. *(A) Fraction of disruptive variants residing on essential genes (pLI ≥ 0.9) or other genes (pLI < 0.9). (B) Probability of haploinsufficiency for genes corresponding to disruptive and non-disruptive alleles. (C) Fraction of disruptive variants residing on gene-encoded proteins with autoinhibitory domains (n = 63) or other genes (n = 1,649). (D) Distribution of s_{het} selection coefficients for genes corresponding to disruptive and non-disruptive alleles. (E) Betweenness centrality values in human interactome for proteins harboring non-disruptive (n = 588), partially disruptive (n = 116), and null-like variants (n = 54). (F) Protein degree in human interactome for proteins harboring non-disruptive (n = 588), partially disruptive (n = 116), and null-like variants (n = 54). Error bars in (A) and (C) indicate +SE of proportion. P values in (A) and (C) by one-tailed Z-test. P values in (B), (D), (E) and (F) by one-tailed U-test. Significant P values in* ***bold****.*

Many other potentially disruptive mutations may be present in more essential genes if other type of mutations, for example de novo, are examined; however, such variants are not characteristic of the ExAC variants sampled in this study.

Topological properties of interactome networks can provide key insights towards the underlying properties of proteins comprising the network. Of particular interest are nodes with high betweenness centrality because such nodes control information flow within a network and correspond well with essential genes (Yu et al., 2007). Therefore, we examined how disruptive variants vary with respect to betweenness centrality and found that null-like variants (**Figure 3.4-2D**) corresponded with proteins with significantly lower betweenness centrality in comparison to partially disruptive variants ($P = 4.7\times10^{-4}$ by one-tailed *U*-test) while no significant difference was observed between partially disruptive and non-disruptive alleles ($P = 0.37$ by one-tailed *U*-test, **Figure 3.4-5E**). Moreover, proteins harboring null-like variants have a significantly lower network degree in comparison to proteins harboring partially disruptive variants ($P = 0.021$ by one-tailed *U*-test) while also showing no difference between partially disruptive and non-disruptive variants ($P = 0.45$ by one-tailed *U*-test, **Figure 3.4-5F**). In principle, null-like mutations disrupt all corresponding protein interactions, nullifying the function of the mutated protein. As such, null-like variants largely accumulate on gene-encoded proteins with minimal influence on protein interaction network topology, reinforcing our finding that deleterious mutations propagate primarily in non-essential genes. While single mutations in less essential regions of human genomes may only marginally impact organism fitness, the accumulation of disruptive mutations across a genome does

influence the genetic background of an individual. As such, studying the impact of disruptive population variants may provide valuable insight towards how marginally impactful variants can influence an individual's predisposing risk to disease.

3.4.6 Structural information details contrasting mechanisms of protein interaction perturbations

Residues located at the interface between two proteins are crucial towards inducing protein-protein interactions (Jones and Thornton, 1996). As such, disruptive population variants within our SNV-interaction network are expected to be enriched at protein interaction interfaces. In agreement, disruptive alleles were found to be substantially enriched at interaction interfaces in comparison to away from the interface (19.5% and 4.1%, respectively; $P = 3.3\times10^{-5}$ by one-tailed *Z*-test, **Figure 3.4-6A**, Methods). However, the occurrence of an allele at an interaction interface is not in itself sufficient for inducing a disruption. Rather, at interaction interfaces, disruptive alleles were often the product of more dissimilar amino acid substitutions in comparison to non-disruptive variants (Median = 81 and 56, respectively; $P = 6.8\times10^{-4}$ by one-tailed *U*-test, **Supplementary Figure C.1-6A**). Among interaction interface variants, we also found that disruptive alleles resulted from more significant changes in binding free energy than non-disruptive alleles ($P = 0.0091$ by one-tailed *U*-test, **Figure 3.4-6B**). Hence, our results affirm that dissimilar amino acid substitutions at interaction interfaces are strong precursors for protein interaction disruptions and should be prioritized when searching for disruptive variants.

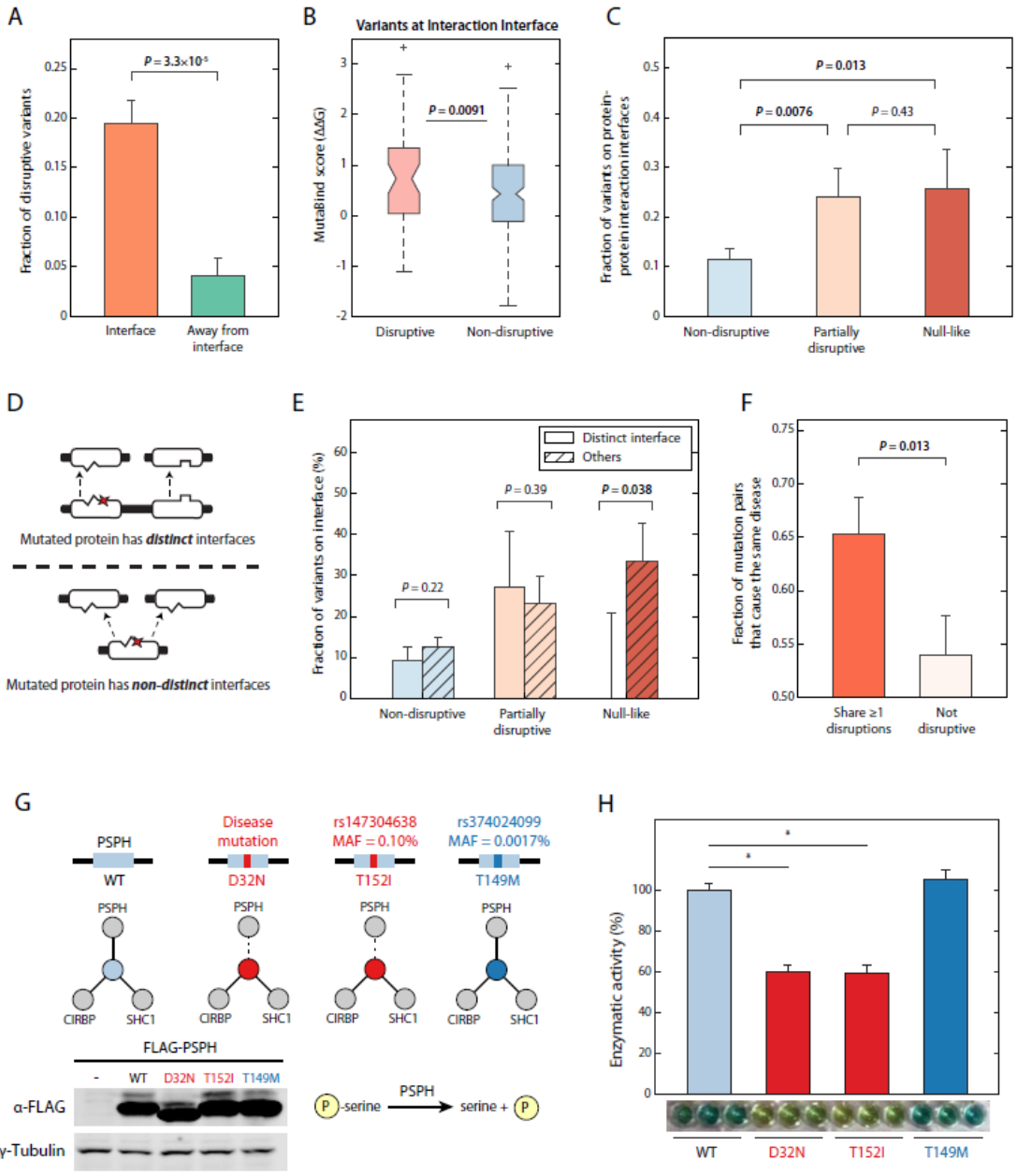


Figure 3.4-6 Identifying candidate disease-associated mutations through shared interaction perturbation profiles. *(A) Fraction of disruptive variants that occur on interaction interfaces in interface domains in comparison to away from interaction interfaces. (B) ΔΔG measured by MutaBind for disruptive and non-disruptive variants on interaction interfaces for interactions with available co-crystal structures in PDB. (C) Fraction of variants on interaction interfaces for interactions with available co-crystal structures in PDB or homology models for non-disruptive (n = 267), partially disruptive (n = 54), or null-like variants (n = 31). (D) Schematic of mutated proteins with distinct or non-distinct interaction interfaces. (E) Fraction*

*of variants on distinct or non-distinct interaction interfaces for non-disruptive (distinct n = 76; non-distinct n = 191), partially disruptive (distinct n = 11; non-distinct n = 43), and null-like (distinct n = 7; non-distinct n = 24) variants. (F) Fraction of mutations pairs that lead to the same disease for germline mutations that share one or more disrupted interactions in common in comparison to pairs that do not disrupt interactions. (G) Schematic of interaction perturbation profiles for disease-associated mutation D32N and rare variants T152I and T149M. Stable expression of FLAG-tagged wild-type and mutant PSPH proteins were validated by Western blot using α-FLAG. α-γ-Tubulin was used as a loading control. A brief diagram of PSPH phosphatase activity is shown. (H) Enzymatic activity of purified recombinant wild-type and mutant PSPH using phosphoserine substrate was measured in vitro using a malachite green assay performed in triplicate. Mutant PSPH enzymatic activity is shown in proportion to wild-type activity. Error bars indicate +SE of mean. * P < 0.01. Error bars in (A), (C), (E), and (F) indicate +SE of proportion. P values in (A), (C), (E), and (F) by one-tailed Z-test. P value in (B) by one-tailed U-test. Significant P values in* ***bold****.*

Mutations on protein interaction interfaces tend to disrupt specific protein interactions while mutations to buried residues are more likely to impact protein stability (Das et al., 2014a; Wang et al., 2012; Zhong et al., 2009). Furthermore, partially disruptive disease-associated mutations have been reported to be strongly enriched on interaction interfaces in comparison to null-like disease-associated mutations (Sahni et al., 2015). To investigate whether partially disruptive and null-like population variants are distributed in a similar manner, we overlaid our tested variants onto their corresponding interaction interfaces using co-crystal structures (Berman et al., 2000) and homology models (Mosca et al., 2013). While partially disruptive variants were found to be enriched on interaction interfaces in comparison to non-disruptive variants ($P = 0.0076$ by one-tailed *U*-test), unexpectedly, we also found that null-like variants were enriched on interaction interfaces relative to non-disruptive variants ($P = 0.013$ by one-tailed *U*-test; **Figure 3.4-6C**). This contrasting result

suggests that null-like variants located on interaction interfaces do not perturb their interaction partners by destabilizing protein folding. This result, though, should only occur if the harboring protein uses a single, shared interaction interface to interact with each of its multiple interaction partners.

To explore whether null-like variants located at interaction interfaces disrupt their multiple interaction partners through a shared interface, we partitioned tested proteins into two sets: (1) proteins with *distinct* interaction interfaces in which not a single residue is shared between at least two different interaction interfaces and (2) proteins with *non-distinct* interaction interfaces in which a subset of residues are shared between all interaction interfaces or only a single interaction interface is available. Upon partitioning proteins in this manner, we found that sampled variants were equally likely to reside on distinct or non-distinct interaction interfaces for both non-disruptive and partially disruptive variants (**Figure 3.4-6D**). In stark contrast, though, null-like variants only resided on proteins with non-distinct interaction interfaces ($P = 0.038$ by one-tailed Z-test; **Figure 3.4-6D**). Hence, null-like population variants in our SNV-interaction network are mostly the consequence of a shared interaction interface disrupting each of the relatively limited number of interactions available to null-like variants (**Figure 3.4-5F**).

Considering that null-like variants are also less likely to be essential (**Figure 3.4-5E**), most disruptive mutations to proteins with multiple interaction partners will be partially disruptive (**Figure 3.4-2D**). Hence, the partially disruptive population variants identified in our results may actually have a larger influence on individual fitness and genetic background since partially disruptive variants are more likely to

occur on essential genes relative to null-like variants. Partially disruptive variants should therefore be carefully considered when searching for potential trait- or disease-associated mutations.

3.4.7 Variants with matching interaction disruption profiles have corresponding molecular phenotypes

Ascertaining whether a disruptive mutation can impact fitness is a challenging task. Nonetheless, disruptions induced by disease-associated mutations are a powerful resource for such a task because the phenotypic impact of such mutations is known and the molecular impact of the disease-associated mutations – through protein interactions and functional assays – is measureable. Hence, we can overlay known disease-associated mutations onto the human interactome and qualitatively inspect where variant-induced disruptions overlap with disease mutations (**Supplementary Figure C.1-6B**). Cases of overlapping disrupted interactions between population variants and known disease-associated mutations are particularly interesting because such overlaps imply that an equally disruptive population variant on the same gene could also lead to the same disease phenotype.

To better assess whether equally disruptive mutations can lead to the same phenotype, we examined all pairs of disease-associated mutations in our HGMD set of tested mutations that occur on the same gene. We then measured the fraction of mutation pairs that cause the same disease for cases in which mutation pairs either (1) shared one or more disrupted interactions in common or (2) neither mutation was disruptive. We found that pairs of disease-associated mutations that share one or more

disrupted interactions resulted in the same disease significantly more often than cases in which no disrupted interactions were found ($P = 0.013$ by one-tailed Z-test, **Figure 3.4-6F**). This result indicates that shared interaction disruption profiles may be an informative approach towards identifying candidate disease-associated mutations.

To demonstrate that shared disrupted interaction profiles between a population variant and disease-associated mutation are indeed informative, we identified two mutations with shared disruption profiles on the phosophoserine phosphatase, PSPH (**Figure 3.4-6G**): (i) T152I, a rare variant (MAF = 0.10%) in ExAC and (ii) D32N, a disease-associated mutation reported to strongly decrease PSPH enzymatic activity in a compound heterozygous individual with phosphoserine phosphatase deficiency (Veiga-da-Cunha et al., 2003). An additional PSPH non-disruptive rare variant, T149M, was included as a control. Next, we purified recombinant wild-type, D32N, T152I, and T149M PSPH proteins and measured for changes phosphatase activity for mutant PSPH relative to wild-type using a malachite green assay. Our *in vitro* assays revealed that T152I reduced PSPH phosphatase activity to 59.2% ± 4.3% ($P = 0.0010$ by one-tailed t-test), which nearly matched the D32N reduction in activity equal to 60.0% ± 3.3% ($P = 6.6\times10^{-4}$ by one-tailed t-test). In contrast, T149M showed no significant change in enzymatic activity relative to wild-type ($P = 0.19$ by one-tailed t-test, **Figure 3.4-6H**). These findings suggest that T152I could result in the same disease phenotype as D32N; however, individuals homozygous for T152I are not reported in ExAC, nor is clinical data available for this variant.

3.5 Discussion

Disentangling the phenotypic impact of functional missense mutations from benign mutations has proved uniquely challenging (Cassa et al., 2013; Dorfman et al., 2010; Masica et al., 2015; Miosge et al., 2015; Wang et al., 2018). Conventions for determining which missense mutations are functional vary widely (Tennessen et al., 2012; The 1000 Genomes Project Consortium, 2012; The Genome of the Netherlands Consortium, 2014), as do their genome-wide conclusions for the number of functional coding mutations per individual (**Figure 3.4-2C**). Therefore, in absence of a consensus metric for assessing the functional impact of missense mutations, we aimed to directly measure the impact of 1,712 missense population variants from ExAC tested across 4,297 protein-variant interaction pairs.

Towards this end, we identified 309 disruptive variants corresponding to 672 disrupted protein interactions. In line with expectations, we found that disruptive variants occurred more frequently as allele frequency decreased, were enriched on interaction interface residues, and predominately targeted evolutionarily conserved protein residues. However, despite this very strong enrichment for disruptive variants at conserved coding sites **(Figure 3.4-4C**), which suggests sites harboring disruptive variants are functionally important, the potentially deleterious effect of disruptive variants appears to be mitigated at both the gene and interactome level. As such, we observed that disruptive alleles were significantly depleted in Loss of Function-Intolerant genes, occurred on genes less likely to be haploinsufficient, and that null-like variants targeted proteins with lower betweenness centrality and lower degree in the human interactome.

Less intuitively, we also found that partially disruptive and null-like variants occurred equally often on interaction interfaces, disruptive variants were enriched in genomic regions under positive selection, and that 10.2% of common variants (MAF > 5%) were disruptive. Moreover, by combining our measured disruption rates with the site frequency spectrum calculated from ExAC, we determined that 11.2% of missense variants per individual genome are expected to be disruptive. Considering that interaction perturbations are just one way in which mutations can be disruptive, our genome-wide figure for the number of disruptive variants per individual represents only a lower-bound estimate. Genome-wide surveys for other types of mutations may reveal that functional variants, at least at the molecular-level, are even more widespread than suggested here.

Although our results corresponded well with expected trends for conservation and allele frequency, we are in no manner implying that the disruptive population variants reported here strongly or even moderately reduce organism fitness. Essential genes, including those associated with autosomal-dominant disease (**Supplementary Figure 6.1-5A**) and known oncogenes (**Supplementary Figure 6.1-5B**), are too infrequently targeted by disruptive variants to suggest otherwise. Nonetheless, the cumulative impact of protein interaction perturbations help constitute the genetic background of an individual. Therefore, examining the molecular phenotypes of disruptive variants and the genetic context in which these variants occur is important in assessing an individual's risk for a particular disease. For example, we identified a null-like variant, A142T, on the protein AKR7A2 that significantly reduces enzymatic activity relative to wild-type (**Figure 3.4-3F**) and segregates at common allele

frequencies (MAF = 6.4%). While evidence exists demonstrating that AKR7A2_A142T lowers drug metabolism (Bains et al., 2010), this variant alone likely has a minimal fitness impact. Disruptive mutations to enzymes in the same pathway as AKR7A2 co-occurring with A142T, though, could potentially compromise the neurotransmitter degradation pathway to which this protein belongs. Indeed, mutations to enzymes in this pathway, ABAT and SSADH, result in severe neurological disorders (Akaboshi et al., 2003; Medina-Kauwe et al., 1998; Tsuji et al., 2010). Hence, we anticipate that the results of our interactome will provide contextual information for many disease-associated mutations, particularly for partial penetrant mutations in which only certain fraction of individuals carrying a mutation are afflicted with disease.

Similar but more direct relationships between disruptive variants and disease can be found in our screen as well. For instance, we identified a population variant on PSPH, T152I (MAF = 0.10%), that reduces enzymatic activity to the same extent as a compound heterozygous mutation, D32N (Veiga-da-Cunha et al., 2003). Therefore, we postulate that T152I could potentially phenocopy D32N in a matching genetic background and elevate risk for phosphoserine phosphatase deficiency in other backgrounds. We further observed that pairs of disease-associated mutations that disrupt the same set of interactions result in the same disease significantly more often than non-disruptive pairs. In addition to T152I and D32N, multiple cases in which a population variant disrupted the same interactions as a disease-associated mutation were found. In this manner, matching disruption profiles can also be used to separate benign variants of unknown significance from those that are disease-associated.

While our high-throughput interaction perturbation approach has allowed us to survey >2,000 SNVs, a computational functional prediction approach is absolutely vital towards achieving the genome-wide scales needed to maintain pace with rapidly proliferating sequencing efforts. Moreover, interaction perturbations are only a strict subset of the variety of ways in which mutations can impair protein function. Approaches assessing other modes of altering protein function, for example, loss of protein-DNA interactions or protein configuration changes, are needed to fully interpret the impact of coding variants. Indeed, all organism-phenotypes must stem from molecular-level perturbations to cellular activity. Therefore, the genetic, protein interaction, and population-level insights presented here could represent a pivotal step forward towards an improved understanding of the evolutionary forces that shape the human genome and protein function.

3.6 Materials and Methods

3.6.1 Selecting single nucleotide variants from ExAC, HGMD, and COSMIC databases

Population variants encoding for missense mutations were selected from ExAC release 1.0. Disease-associated missense mutations were obtained from HGMD public release version 2014. Cancer somatic missense mutations were selected from COSMIC version 75. For all three sets, we required that mutations reside on genes in either hORFeome v8.1 (Yang et al., 2011) or v5.1 (The MGC Project Team, 2009), corresponded with one or more Y2H-testable protein-protein interactions from [19-22], and, for ExAC variants, achieved a PASS filter status. We mapped each RefSeq

transcript from ExAC to an appropriate ORF in our library by looking at the top blastx candidate with an e-value ≤ 0.001. We verified that this is a representative ORF for our mutation by performing EDNAFULL matrix pairwise alignment using EMBOSS Stretcher. Valid representative ORFs must be identical in the 31 amino acid window centered on the position of interest for mutagenesis. Beyond local identity, ORFs are required to have more than 95% global identity, or be an exact subset of the transcript, spanning at least a third of the query transcript.

To minimize gene bias, we selected an average of two variants per gene. Since over half of variants in ExAC are singletons, to avoid oversampling rare alleles, we selected between 200-400 variants across six mutually exclusive allele count bins of 1, <10, <100, <1,000, < 10,000, and >10,000 for a total of 1,712 ExAC alleles (**Figure 3.4-1B**). 204 HGMD mutations were selected in accordance to criteria detailed in (Wei et al., 2014) but expanded to test across all amenable Y2H protein-protein interactions. 169 COSMIC mutations were also tested among 116 different genes with available hORFeome clones for testing across all amenable Y2H protein-protein interactions. Genes listed in Cancer Gene Census (v75) and listed as a known driver in IntOGen (2014.12) were designated as *Known cancer genes*. Genes not listed in Cancer Gene Census and not listed as a driver in IntOGen were designated as *Other genes* (**Figure 3.4-2B**).

3.6.2 Large-scale cloning of SNVs through Clone-seq pipeline

Single colony-derived mutant clones were constructed using a previously described methodology named Clone-seq (Wei et al., 2014), a high-throughput mutagenesis and

next-generation sequencing platform (**Supplementary Figure C.1-1**). In brief, wild-type clones were picked from hORFeome clones and served as templates for site-directed mutagenesis performed at 96-well scales using site-specific mutagenesis primers (Eurofins). To minimize sequencing artifacts, PCR was limited to 18 cycles using Phusion polymerase (NEB, M0530). PCR product was digested overnight with DpnI (NEB, R0176) then transformed into competent bacteria cells to isolate single colonies. Up to four colonies per individual mutagenesis reaction were then hand-picked and arrayed into 96-well plates and incubated for 21 hrs at 37°C under constant vibration. After incubation, glycerol stocks were generated then clones were pooled into independent bacterial pools. An additional maxiprepped bacterial pool consisting of only wild-type DNA templates corresponding to each mutagenesis PCR reaction was also prepared. Maxiprepped clonal DNA from each bacterial pool were then combined through multiplexing (NEB, E7335) and sequenced in a single 1x75 single-end Illumina NextSeq run. Properly mutated clones which differed from their sequenced wild-type templates only by the desired single base-pair mutation – and nowhere else – were then identified by next-generation sequencing analysis and recovered from their corresponding single colony glycerol stocks.

3.6.3 Identifying successfully mutated clones and filtering clones with unwanted mutations

After de-multiplexing, mapped reads corresponding to the generated pools (wildtype plus up to four mutant pools) were mapped to genes of interest using the BWA *mem* algorithm (`bwa mem -a -t 12 <reference> <reads>`). In order to detect

both the desired variant as wells as undesired off-target mutations, we first obtained the read counts for each allele (A, T, C, G, insertion, or deletion) for all positions in the clones. Using these read counts we calculated the score for a given position, pos, containing a mutation from the wildtype allele, WT, to a mutant allele, Mut, as follows:

$$Score(WT, pos, Mut) = \frac{Observed_{Mut,pos}}{Expected_{Mut,pos}}$$

where $Observed_{Mut,pos}$ is the observed fraction of reads at position `pos` matching allele `Mut` and $Expected_{Mut,pos}$ is the fraction of reads at position `pos` matching allele `Mut` that we would expect to see if it the mutation in question had indeed occurred. We define this fraction as:

$$Expected_{Mut,pos} = \frac{1}{TotalMutations} + (TotalMutations - 1) * SeqErr(pos) - (Alleles - 1) * SeqErr(pos)$$

where $TotalMutations$ is the total number of mutants attempted on a particular ORF (i.e. the number of copies of the ORF included in the pool), $SeqErr(pos)$ refers to the inherent sequencing error, and $Alleles$ is the total number of alleles.

To explain further, assuming that all clones for a particular gene contribute a similar number of reads, we expect that if one of the clones for a gene contains a mutation to the `Mut` allele at position `pos`, we should see $\frac{1}{TotalMutations}$ fraction of the reads match the `Mut` allele. Due to sequencing errors, we expect the true fraction observed to deviate slightly from this base fraction. We first add a term for the fraction of `Mut` alleles that we expect to see as a result of sequencing errors in the other non-mutant clones for the gene. Second, we subtract a term for sequencing errors in the mutant clone converting `Mut` allele to any of the $(Alleles - 1)$ other alleles. We

define the sequencing error as the average fraction of non-WT bases observed in the ten closest positions that were not targeted for mutagenesis.

Based on comparisons to Sanger sequencing results, we set a threshold of $Score(WT, pos, Mut) \geq 0.5$ to call true mutations. In identifying successful instances of site-directed mutagenesis, we first checked for presence of the desired mutation using this score threshold. Using the scores for all other positions along the clone, we then screened each successful mutant for the presence of any other unwanted mutations that may have been introduced as PCR artifacts. Any clones containing unwanted mutations were removed, and the remaining clones were sorted using a combination of their desired mutation score, maximum undesired mutation score, sequencing coverage, and sequencing quality.

3.6.4 Profiling disrupted protein-protein interactions by high-throughput Y2H

Clone-seq identified mutant clones were transferred into Y2H vectors pDEST-AD and pDEST-DB by Gateway LR reactions then transformed into *MAT***a** Y8800 and *MAT*α Y8930, respectively. All DB-ORF *MAT*α transformants, including wild-type ORFs, were then mated against corresponding wild-type (WT) and mutant AD-ORF *MAT***a** transformants in a pairwise orientation using automated 96-well procedures to inoculate AD-ORF and DB-ORF yeast cultures followed by mating on YEPD agar plates. All DB-ORF yeast cultures were also mated against *MAT***a** yeast transformed with empty pDEST-AD vector to screen for autoactivators. After overnight incubation at 30°C, yeast were replica-plated onto selective Synthetic Complete agar media lacking leucine and tryptophan (SC-Leu-Trp) to select for mated, diploid yeast then

incubated again overnight at 30°C. Diploid yeast were then replica-plated onto SC-Leu-Trp agar plates also lacking histidine and supplemented with 1 mM of 3-amino-1,2,4-triazole (SC-Leu-Trp-His+3AT) as well as SC-Leu-Trp agar plates lacking adenine (SC-Leu-Trp-Ade). After overnight incubation at 30°C, plates were replica-cleaned and incubated again for three days at 30°C.

Disrupted protein-protein interactions were identified as follows: (1) mutated protein reduces growth by at least 50% relative to wild-type interaction as benchmarked by twofold serial dilution experiments, (2) neither wild-type or mutant DB-ORFs are autoactivators, (3) reduced growth phenotype reproduces across three screens. A mutation was scored as disruptive if one or more corresponding protein-protein interactions were disrupted and scored as non-disruptive otherwise.

3.6.5 Assessing genome-wide functional mutation rates for coding variants

The total number of missense variants in ExAC listed in Figure 3.4-1C was determined by summing the adjusted overall allele count found in the ExAC database for all variants annotated as missense_variant in at least one transcript. The number of functional mutations was calculated by multiplying the mean disruption rate per individual (**Appendix C.2.1**) by the total number of missense variants in ExAC.

The total number of missense variants in the 1000 Genomes Consortium – Phase I, Genomes of the Netherlands, and Exome Sequencing Project – Phase I were obtained from (The 1000 Genomes Project Consortium, 2012), (The Genome of the Netherlands Consortium, 2014), and (Tennessen et al., 2012), respectively. Calculations for the number of functional missense mutations from each source are

annotated in Supplementary Tables C.3-1, C.3-2, and C.3-3. We note that the number of functional mutations by mutation type was not reported for ESP variants in (Tennessen et al., 2012). As such, functional nonsynonymous mutations, including nonsense variants, were instead reported for ESP – Phase I. We expect the percent range of functional missense variants for ESP, 5.5% and 10.0% (**Supplementary Table C.3-3**), to be small overestimates as a result.

3.6.6 Orthogonal validation of disrupted and non-disrupted interactions by Protein Complementation Assay

To validate that variant-disrupted protein-protein interactions are reproducible across a different assay, we systemically selected a subset of Y2H-tested interactions for retest by PCA. Bait ORFs in pDONR223 for disruptive and non-disruptive variants were transferred into F1 Venus fragments while prey ORFs for corresponding interaction partners were transferred into to F2 Venus fragments using Gateway LR reactions for a total of 408 ORF-interaction pairs comprising 401 gene-level interactions. Bait and prey ORF pairs were then randomly scrambled across 92 PRS and 92 RRS ORF pairs previously described in (Braun et al., 2009; Venkatesan et al., 2009) to minimize detection bias across different 96-well plates.

To perform PCA, HEK293T cells were seeded onto black 96-well flat-bottom dishes (Costar, 3603). HEK293T cells were maintained in complete DMEM medium supplemented with 10% FBS and incubated at 37°C under air with 5% CO_2. Cells were grown to 60-70% confluency then co-transfected using 100 ng of bait vector plus 100 ng of prey vector with 1.0 μL of 1 mg/mL PEI (Polysciences Inc, 23966) mixed

thoroughly with 20 µL OptiMEM (Gibco, 31985-062) per transfection. After 72 hrs incubation at 37°C, a Tecan M1000 plate reader was used to measure PCA fluorescence (excitation = 514 nm; excitation = 527 nm) for all samples. A manually-adjusted gain was applied to ensure all measurements were performed within a linear range. Detection thresholds were selected such that ORF pairs resulting in a signal greater than the threshold were scored as "detected" while scores that fell below the threshold were "undetected." The fraction of recovered pairs represents the proportion of ORF pairs that scored above a given threshold over the total set of ORF pairs tested per category. As a quality control measure, interaction pairs in which either a bait or prey ORF did not amplify by PCR using F1 Venus- or F2 Venus-specific primers, respectively, were removed from PCA detection rate calculations.

3.6.7 Construction of vectors for dual-fluorescent screen and Western blot

Gateway LR reactions were used to transfer ORFs into mammalian expression vectors. pDEST-DUAL vector for dual-fluorescence screen was constructed by inserting an mCherry cassette independently driven by a minCMV promoter into pcDNA-DEST47 (Invitrogen, 12281-010), which features a C-terminal GFP tag. PSPH wild-type, D32N, T152I, and T149M were transferred into a pQXIP (ClonTech, 631516) vector modified to include a Gateway cassette featuring a C-terminal 3xFLAG. AKR7A2 wild-type and A142T were transferred into pcDNA-DEST40, which includes a V5 tag (Invitrogen, 12274-015).

3.6.8 Dual-fluorescence assay to measure impact of variants on protein stability

In order to screen for variants that destabilize protein expression, we first screened for

stably expressed GFP-tagged wild-type proteins. To do this, we transferred wild-type ORFs into pDEST-DUAL by Gateway LR reactions. HEK293T cells were then seeded onto black 96-well flat-bottom dishes (Costar, 3603) HEK293T cells were maintained in complete DMEM medium supplemented with 10% FBS. All cell incubation steps were performed at 37°C under air with 5% CO_2. Cells were grown to 60% confluency then co-transfected using 150 ng of sample DNA in pDEST-DUAL and 1.0 µL of 1 mg/mL PEI (Polysciences Inc, 23966) mixed thoroughly with 20 µL OptiMEM (Gibco, 31985-062). Four replicates of empty pDEST-DUAL and four replicates of empty pcDNA-DEST47 were also transfected per 96-well plate as positive controls for mCherry expression and negative controls for GFP expression, respectively. After 72 hrs incubation, stably expressed wild-type GFP-tagged proteins were identified using a Tecan M1000 plate reader. Samples that resulted in GFP and mCherry expression significantly above background were then validated by automated fluorescence microscopy using an ImageXpress system. In this manner, we identified 202 wild-type genes corresponding to 291 ExAC variants. Single clones for ExAC variants were then transferred into pDEST-DUAL by Gateway LR reactions for further screening.

Wild-type and mutant ORF pairs in pDEST-DUAL were transfected into 293T in the same manner as described for our first wild-type screen, including the same eight pDEST-DUAL and pcDNA-DEST47 controls per plate. Mutant ORFs corresponding to a particular wild-type ORF were always partitioned onto the same plate. After 72 hrs incubation, GFP and mCherry fluorescence readings using a Tecan M1000 plate reader were measured for all samples and processed into stable,

moderately stable and unstable categories as described in Appendix C.2.2. Samples were also imaged automated fluorescence microscopy using an ImageXpress system.

3.6.9 Cell culture for Western blotting

HEK293T cells were maintained in complete DMEM medium supplemented with 10% FBS and incubated at 37°C under air with 5% CO_2. Cells were grown in 6-well dishes to 70-80% confluency then transfected using 2 μg of vector with 10 μL of 1mg/mL PEI (Polysciences Inc, 23966) mixed thoroughly with 150 μL OptiMEM (Gibco, 31985-062). After 24 hrs incubation, cells were gently washed three times in 1x PBS and then resuspended in 200 μL cell lysis buffer [10 mM Tris-Cl pH 8.0, 137mM NaCl, 1% Triton X-100, 10% glycerol, 2 mM EDTA, and 1x EDTA-free Complete Protease Inhibitor tablet (Roche)] and incubated on ice for 30 min. Extracts were cleared by centrifugation for 10 mins at 13,000 rpm at 4℃. Samples were then treated in 6x SDS protein loading buffer (10% SDS, 1 M Tris-Cl pH 6.8, 50% glycerol, 10% β-mercaptoethanol, 0.03% Bromophenol blue) and subjected to SDS-PAGE. Proteins were then transferred from gels onto PVDF (Amersham) membranes. Anti-FLAG (Sigma, F1804), anti-V5 (Invitrogen, R960-25), anti-GFP (SCBT, 9996) and anti-γ-Tubulin (Sigma, T5192) at 1:5000, 1:3000, 1:1000, and 1:3000 dilutions, respectively, were used for immunoblotting analysis.

3.6.10 Protein purification of recombinant PSPH and AKR7A2

Gene-specific primers were used to clone BamHI and XhoI restriction endonuclease digestion sites onto the 5’ and 3’ ends of ORFs for wild-type, D32N, T152I, and T149M clones of PSPH by PCR. PCR products as well as a pET28a-based, custom

modified pET-6xHis-SUMO expression vector were then digested overnight using BamHI (NEB, R3136) and Xho1 (NEB, R0146) restriction endonucleases. All digested products were cleaned up by gel extraction. PCR products were then ligated into cut pET-6xHis-SUMO vector by 10 μL T4 ligase (NEB, M0202) reactions using a 3:1 ratio of insert to template incubated for 30 min at RT. Ligated products were then transformed into competent cells and plated to isolate single colonies. Properly ligated colonies were validated by colony PCR. Colony PCR-validated pET-6xHis-SUMO PSPH constructs were then transformed into Rosetta strain competent bacteria cells (Novagen, 71401-3).

To purify recombinant wild-type and mutant PSPH protein, single colonies of transformed Rosetta strain bacteria were inoculated overnight for use as starter cultures. Starter cultures were then used to inoculate 1.0 L LB media including kanamycin and chloramphenicol and grown 2-4 hrs at 37°C, shaking at 250 rpm until OD600 = 0.6. 200 μL of 1 M IPTG was then added to induce protein expression. Induced cultures were then incubated for 18 hrs at 18°C, shaking at 250 rpm. After incubation, cultures were then centrifuged at 4,000xg for 20 min at 4°C. Supernatant was discarded and pellet was resuspended in 35 mL Resuspension Buffer (500mM NaCL, 50mM Tris-base pH8.0) on ice. Note that, unless stated otherwise, all steps moving forward were performed on ice or at 4°C. Resuspended pellet was then sonicated to lyse cells and then centrifuged at 16,000xg for 45 min. Supernatant was then run through a column prewashed with Wash Buffer (20mM NaCl, 20mM Tris pH7.5) and loaded with Cobalt agarose beads (GoldBio, H-310) for purification of 6x His-tagged protein. Purified samples bound to Cobalt beads were then treated

overnight with lab-purified Ulp1 protease for SUMO tag cleavage. Afterwards, samples were again run through a column prewashed with Resuspension Buffer and eluted sample was collected. Lastly, purified protein samples were fractionated by FPLC and samples lacking detectable SUMO expression by Coomassie gel were used for experiments.

Wild-type and mutant A142T were prepared in the same manner as PSPH except for the following changes: (1) AKR7A2 gene-specific primers were used for PCR, followed by EcoRI (NEB, R3101) and Xho1 (NEB, R0146) double digestion of PCR product and pET-6xHis-SUMO vector; (2) after induction with 200 μL of 1 M IPTG, cultures were incubated for 5 hrs at 37°C, shaking at 250 rpm.

3.6.11 Phosphatase activity measurements for PSPH variants

Wild-type and mutant PSPH activity were measured using a malachite green assay as follows: Malachite Green Reagent Stock was prepared by combining 30 mL Malachite Green (Sigma, M9636) with 20 mL 4.2% ammonium molybdate (Sigma, 277908) / 4M HCL and mixing for > 30 min. Malachite Green Reagent Stock was filtered through a 0.2 μm filter unit and stored at 4°C. Malachite Green Working Reagent was then prepared by adding Tween-20 to a final concentration of 0.01% in Malachite Green Reagent Stock. Using a 96-well plate (Costar, 3696), A_{620} for sodium phosphate in Malachite Green Working Reagent at concentrations of 10, 15, 20, 25, 30, 35, and 40 μM at pH 7.4 was then measured using a Tecan M1000 plate reader to generate a standard curve. Next, 100 ng of purified recombinant PSPH protein was then added to 20 μL total of Assay Buffer (30 mM HEPES at pH 7.4, 1 mM EGTA, 1

mM $MgCl_2$ and 100 μM phosphoserine) and mixed with 80 μL Malachite Green. Negative controls lacking recombinant protein or phosphoserine substrate were also included. After plate incubation at 37°C for 5 min, A_{620} was measured for all samples. Percent change in phosphatase activity for mutant PSPH was measured as the ratio of mean mutant PSPH activity to mean wild-type PSPH enzymatic activity over three replicates.

3.6.12 Kinematic measurement of SSA turnover by AKR7A2

Using a 96-well plate (Costar, 3696), A_{340} for NADPH (Cayman, 9000743) at concentrations of 2000, 1000, 500, 250, 125, 62.5, 31.3, 15.6, and 0 μM in 100 μM sodium phosphate buffer at pH 8.0 was then measured using a Tecan M1000 plate reader to generate a standard curve. To measure NADPH-dependent turnover of succinic semialdehyde (SSA) to γ-Hydroxybutyrate (GHB) for wild-type AKR7A2 and mutant A142T AKR7A2, 3.0 μg of purified AKR7A2 protein was added to SSA aliquoted to individual wells in 96-well plate at concentrations of 0.50, 0.75, 1.0, 1.5, 2.0, 2.5, 3.0, and 3.5 mM in 100 μM sodium phosphate buffer at pH 8.0. Reactions were started simultaneously by adding in NADPH at an initial concentration of 0.5 mM and incubated at 37°C. Negative controls lacking recombinant protein, NADPH, or SSA were also included. OD320 measurements were taken every 60 seconds for a total of 15 minutes. AKR7A2 wild-type and A142T experiments were performed over three replicates.

3.6.13 Enrichment of disruptive mutations on interaction interfaces

We examined the loci of ExAC variant residues relative to protein-protein interaction

interfaces. "On interface" was defined as either at an interface residue or in the interface domain, while "away from interface" was defined as neither at an interface residue nor in the interface domain. Interface residues and domains were defined as previously described in IntASA (Das et al., 2014b) and Instruct (Meyer et al., 2013). Interface residues in context of non-disruptive, partially disruptive, and null-variants were acquired from the Interactome INSIDER (Meyer et al., 2018) database for structures annotated as PDB or I3D. Interface residues for interactions with available PDB co-crystal structures not annotated in Interactome INSIDER were determined as previously described in (Meyer et al., 2018).

3.6.14 Metrics for evolutionary site conservation and ancestral alleles

Ancestral alleles were parsed from dbSNP build 150. Only sites in which the ancestral allele matched the corresponding major or minor allele tested in our SNV-interaction network were used for calculations. Jensen-Shannon divergence (JSD) scores were obtained as previously described in (Meyer et al., 2018). To measure the average overall allele frequency across different JSD scores, cutoff scores of 0.2, 0.3, … , 1.0 were applied and the overall allele frequencies per tested ExAC variant were averaged cumulatively across each cutoff score. phyloP scores were obtained using the Table Browser of the UCSC Genome Browser and inputting the hg19 coordinates for each tested variant.

3.6.15 Signals of positive selection for disruptive alleles

Fay and Wu's H was applied genome-wide across 3 kb sliding windows using the 1000 Genomes Phase 3 dataset (The 1000 Genomes Project Consortium, 2015).

Regions with a Fay and Wu's H test statistic at or above the 95th percentile were considered significant and were calculated for the overall population as well as in AFR, EUR, EAS, and SAS populations individually. The number of disruptive variants that occur in regions with a significant Fay and Wu's H test statistic were summed over all variants that occur in any region with a measurable Fay and Wu's H statistic.

3.6.16 Protein interaction network-based calculations of betweenness centrality and degree

The human reference interactome was assembled as described in (Meyer et al., 2018). Using this assembled human interactome, betweenness centrality and degree for proteins corresponding to tested ExAC variants were calculated using the *betweenness_centrality* and *degree* functions of NetworkX.

3.6.17 Curation of inheritance patterns and phenotypes for disease-associated genes

Disease-associated genes with autosomal dominant inheritance patterns were obtained from (Singh et al., 2014) and cross validated against OMIM (accessed 10/31/2017). Genes exclusively annotated in (Singh et al., 2014) as autosomal dominant were compared against all other genes listed in (Singh et al., 2014). Disease phenotypes for mutations pairs that map back to the same or different disease were obtained from HGMD release version 2017. Mutation pairs were deemed to cause the same disease if strings for their corresponding disease phenotypes were equal or both have matching UMLS mappings.

3.7 References

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat Meth *7*, 248-249.

Akaboshi, S., Hogema, B.M., Novelletto, A., Malaspina, P., Salomons, G.S., Maropoulos, G.D., Jakobs, C., Grompe, M., and Gibson, K.M. (2003). Mutational spectrum of the succinate semialdehyde dehydrogenase (ALDH5A1) gene and functional analysis of 27 novel disease-causing mutations in patients with SSADH deficiency. Human Mutation *22*, 442-450.

Asthana, S., Roytberg, M., Stamatoyannopoulos, J., and Sunyaev, S. (2007). Analysis of Sequence Conservation at Nucleotide Resolution. PLOS Computational Biology *3*, e254.

Bains, O.S., Grigliatti, T.A., Reid, R.E., and Riggs, K.W. (2010). Naturally Occurring Variants of Human Aldo-Keto Reductases with Reduced In Vitro Metabolism of Daunorubicin and Doxorubicin. Journal of Pharmacology and Experimental Therapeutics *335*, 533.

Bartha, I., di Iulio, J., Venter, J.C., and Telenti, A. (2017). Human gene essentiality. Nature Reviews Genetics *19*, 51.

Bartha, I., Rausell, A., McLaren, P.J., Mohammadi, P., Tardaguila, M., Chaturvedi, N., Fellay, J., and Telenti, A. (2015). The Characteristics of Heterozygous Protein Truncating Variants in the Human Genome. PLOS Computational Biology *11*, e1004647.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Research *28*, 235-242.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., and Rubin, E.M. (2003). Phylogenetic Shadowing of Primate Sequences to Find Functional Regions of the Human Genome. Science *299*, 1391-1394.

Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., Karchin, R., Kinzler, K.W., Vogelstein, B., and Nowak, M.A. (2010). Accumulation of driver and passenger mutations during tumor progression. Proceedings of the National Academy of Sciences *107*, 18545.

Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J.M., Murray, R.R., Roncari, L., de Smet, A.-S*., et al.* (2009). An experimentally derived confidence score for binary protein-protein interactions. Nat Meth *6*, 91-97.

Cassa, C.A., Tong, M.Y., and Jordan, D.M. (2013). Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. Human mutation *34*, 1216-1220.

Chow, C.Y. (2015). Bringing genetic background into focus. Nature Reviews Genetics *17*, 63.

Cooper, G.M., Brudno, M., Program, N.C.S., Green, E.D., Batzoglou, S., and Sidow, A. (2003). Quantitative Estimates of Sequence Divergence for Comparative Analyses of Mammalian Genomes. Genome Research *13*, 813-820.

Corder, E., Saunders, A., Strittmatter, W., Schmechel, D., Gaskell, P., Small, G., Roses, A., Haines, J., and Pericak-Vance, M. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science *261*, 921-923.

Das, J., Fragoza, R., Lee, H.R., Cordero, N.A., Guo, Y., Meyer, M.J., Vo, T.V., Wang, X., and Yu, H. (2014a). Exploring mechanisms of human disease through structurally resolved protein interactome networks. Molecular BioSystems *10*, 9-17.

Das, J., Lee, H.R., Sagar, A., Fragoza, R., Liang, J., Wei, X., Wang, X., Mort, M., Stenson, P.D., Cooper, D.N.*, et al.* (2014b). Elucidating Common Structural Features of Human Pathogenic Variations Using Large-Scale Atomic-Resolution Protein Networks. Human Mutation *35*, 585-593.

Deary, I.J., Whiteman, M.C., Pattie, A., Starr, J.M., Hayward, C., Wright, A.F., Carothers, A., and Whalley, L.J. (2002). Cognitive change and the APOE ε4 allele. Nature *418*, 932.

Dorfman, R., Nalpathamkalam, T., Taylor, C., Gonska, T., Keenan, K., Yuan, X.W., Corey, M., Tsui, L.C., Zielenski, J., and Durie, P. (2010). Do common in silico tools predict the clinical consequences of amino-acid substitutions in the CFTR gene? Clinical Genetics *77*, 464-473.

Florez, J.C., Jablonski, K.A., Sun, M.W., Bayley, N., Kahn, S.E., Shamoon, H., Hamman, R.F., Knowler, W.C., Nathan, D.M., and Altshuler, D. (2007). Effects of the type 2 diabetes-associated PPARG P12A polymorphism on progression to diabetes and response to troglitazone. The Journal of clinical endocrinology and metabolism *92*, 1502-1509.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A.*, et al.* (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Research *39*, D945-D950.

Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Altshuler, D., Shendure, J., Nickerson, D.A.*, et al.* (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216-220.

Gibson, G. (2011). Rare and Common Variants: Twenty arguments. Nature reviews Genetics *13*, 135-145.

Gorlov, I.P., Gorlova, O.Y., Sunyaev, S.R., Spitz, M.R., and Amos, C.I. (2008). Shifting Paradigm of Association Studies: Value of Rare Single-Nucleotide Polymorphisms. American Journal of Human Genetics *82*, 100-112.

Grantham, R. (1974). Amino Acid Difference Formula to Help Explain Protein Evolution. Science *185*, 862.

Guharoy, M., and Chakrabarti, P. (2005). Conservation and relative importance of residues across protein-protein interfaces. Proceedings of the National Academy of Sciences *102*, 15447-15452.

Henn, B.M., Botigué, L.R., Bustamante, C.D., Clark, A.G., and Gravel, S. (2015). Estimating Mutation Load in Human Genomes. Nature reviews Genetics *16*, 333-343.

Jones, S., and Thornton, J.M. (1996). Principles of protein-protein interactions. Proceedings of the National Academy of Sciences *93*, 13-20.

Keinan, A., and Clark, A.G. (2012). Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. Science *336*, 740-743.

Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A.*, et al.* (2013). Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. Science *342*.

Kim, P.M., Lu, L.J., Xia, Y., and Gerstein, M.B. (2006). Relating Three-Dimensional Structures to Protein Networks Provides Evolutionary Insights. Science *314*, 1938-1941.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B.*, et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285-291.

Lockless, S.W., and Ranganathan, R. (1999). Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. Science *286*, 295-299.

Lyon, R.C., Johnston, S.M., Watson, D.G., McGarvie, G., and Ellis, E.M. (2007). Synthesis and Catabolism of γ-Hydroxybutyrate in SH-SY5Y Human Neuroblastoma Cells: Role of the Aldo-Keto Reductase AKR7A2. Journal of Biological Chemistry *282*, 25986-25992.

Maher, M.C., Uricchio, L.H., Torgerson, D.G., and Hernandez, R.D. (2012). Population genetics of rare variants and complex diseases. Human heredity *74*, 118-128.

Manolio, T.A., Brooks, L.D., and Collins, F.S. (2008). A HapMap harvest of insights into the genetics of common disease. The Journal of Clinical Investigation *118*, 1590-1605.

Margulies, E.H., Blanchette, M., Program, N.C.S., Haussler, D., and Green, E.D. (2003). Identification and Characterization of Multi-Species Conserved Sequences. Genome Research *13*, 2507-2518.

Marth, G.T., Yu, F., Indap, A.R., Garimella, K., Gravel, S., Leong, W.F., Tyler-Smith, C., Bainbridge, M., Blackwell, T., Zheng-Bradley, X.*, et al.* (2011). The functional spectrum of low-frequency coding variation. Genome Biol *12*, R84.

Masica, D.L., Li, S., Douville, C., Manola, J., Ferris, R.L., Burtness, B., Forastiere, A.A., Koch, W.M., Chung, C.H., and Karchin, R. (2015). Predicting survival in head and neck squamous cell carcinoma from TP53 mutation. Human genetics *134*, 497-507.

McDonald, J.H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature *351*, 652.

Medina-Kauwe, L.K., Nyhan, W.L., Gibson, K.M., and Tobin, A.J. (1998). Identification of a Familial Mutation Associated with GABA-Transaminase Deficiency Disease. Neurobiology of Disease *5*, 89-96.

Meyer, M.J., Beltrán, J.F., Liang, S., Fragoza, R., Rumack, A., Liang, J., Wei, X., and Yu, H. (2018). Interactome INSIDER: a structural interactome browser for genomic studies. Nature Methods *15*, 107.

Meyer, M.J., Das, J., Wang, X., and Yu, H. (2013). INstruct: a database of high-quality 3D structurally resolved protein interactome networks. Bioinformatics.

Mintseris, J., and Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein–protein interactions. Proceedings of the National Academy of Sciences of the United States of America *102*, 10930-10935.

Miosge, L.A., Field, M.A., Sontani, Y., Cho, V., Johnson, S., Palkova, A., Balakishnan, B., Liang, R., Zhang, Y., Lyon, S.*, et al.* (2015). Comparison of predicted and actual consequences of missense mutations. Proceedings of the National Academy of Sciences *112*, E5189-E5198.

Mosca, R., Ceol, A., and Aloy, P. (2013). Interactome3D: adding structural details to protein networks. Nat Meth *10*, 47-53.

Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St. Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.-A., Fraser, D.*, et al.* (2012). An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. Science *337*, 100-104.

Online Mendelian Inheritance of Man OMIM McKusick-Nathans Institute of Genetic Medicine (Baltimore, MD).

Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K.A., Lin, G.N., Nam, H.-J., Mort, M., Cooper, D.N., Sebat, J., Iakoucheva, L.M.*, et al.* (2017). MutPred2: inferring the molecular and phenotypic impact of amino acid variants. bioRxiv.

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

Pufall, M.A., and Graves, B.J. (2002). Autoinhibitory Domains: Modular Effectors of Cellular Regulation. Annual Review of Cell and Developmental Biology *18*, 421-462.

Robitaille, J., Després, J.P., Pérusse, L., and Vohl, M.C. (2003). The PPAR-gamma P12A polymorphism modulates the relationship between dietary fat intake and components of the metabolic syndrome: results from the Québec Family Study. Clinical Genetics *63*, 109-116.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, Samuel J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R.*, et al.* (2014). A Proteome-Scale Map of the Human Interactome Network. Cell *159*, 1212-1226.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N.*, et al.* (2005). Towards a proteome-scale map of the human protein-protein interaction network. Nature *437*, 1173-1178.

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J.I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G.I., Wang, Y.*, et al.* (2015). Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders. Cell *161*, 647-660.

Schoenrock, A., Burnside, D., Moteshareie, H., Pitre, S., Hooshyar, M., Green, J.R., Golshani, A., Dehne, F., and Wong, A. (2017). Evolution of protein-protein interaction networks in yeast. PLOS ONE *12*, e0171920.

Singh, Param P., Affeldt, S., Cascone, I., Selimoglu, R., Camonis, J., and Isambert, H. On the Expansion of "Dangerous" Gene Repertoires by Whole-Genome Duplications in Early Vertebrates. Cell Reports *2*, 1387-1398.

Singh, P.P., Affeldt, S., Malaguti, G., and Isambert, H. (2014). Human Dominant Disease Genes Are Enriched in Paralogs Originating from Whole Genome Duplication. PLoS Computational Biology *10*, e1003754.

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

Strittmatter, W.J., Saunders, A.M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G.S., and Roses, A.D. (1993). Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. Proceedings of the National Academy of Sciences of the United States of America *90*, 1977-1981.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G.*, et al.* (2012). Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science *337*, 64-69.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68.

The Genome of the Netherlands Consortium (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet *46*, 818-825.

The MGC Project Team (2009). The completion of the Mammalian Gene Collection (MGC). Genome Research *19*, 2324-2333.

The UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82-90.

Tsuji, M., Aida, N., Obata, T., Tomiyasu, M., Furuya, N., Kurosawa, K., Errami, A., Gibson, K.M., Salomons, G.S., Jakobs, C.*, et al.* (2010). A new case of GABA transaminase deficiency facilitated by proton MR spectroscopy. Journal of Inherited Metabolic Disease *33*, 85-90.

Vallender, E.J., and Lahn, B.T. (2004). Positive selection on the human genome. Human Molecular Genetics *13*, R245-R254.

Veiga-da-Cunha, M., Collet, J.-F., Prieur, B., Jaeken, J., Peeraer, Y., Rabbijns, A., and Van Schaftingen, E. (2003). Mutations responsible for 3-phosphoserine phosphatase deficiency. Eur J Hum Genet *12*, 163-166.

Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.-I*., et al.* (2009). An empirical framework for binary interactome mapping. Nat Meth *6*, 83-90.

Vidal, M. (2001). A Biological Atlas of Functional Maps. Cell *104*, 333-339.

Vidal, M., Cusick, Michael E., and Barabási, A.-L. (2011). Interactome Networks and Human Disease. Cell *144*, 986-998.

Wang, T., Bu, C.H., Hildebrand, S., Jia, G., Siggs, O.M., Lyon, S., Pratt, D., Scott, L., Russell, J., Ludwig, S*., et al.* (2018). Probability of phenotypically detectable protein damage by ENU-induced mutations in the Mutagenetix database. Nature Communications *9*, 441.

Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic Screens in Human Cells Using the CRISPR-Cas9 System. Science *343*, 80.

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotech *30*, 159-164.

Wei, X., Das, J., Fragoza, R., Liang, J., Bastos de Oliveira, F.M., Lee, H.R., Wang, X., Mort, M., Stenson, P.D., Cooper, D.N*., et al.* (2014). A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. PLoS Genet *10*, e1004819.

Yang, X., Boehm, J.S., Yang, X., Salehi-Ashtiani, K., Hao, T., Shen, Y., Lubonja, R., Thomas, S.R., Alkan, O., Bhimdi, T*., et al.* (2011). A public genome-scale lentiviral expression library of human ORFs. Nature methods *8*, 659-661.

Yu, H., Kim, P.M., Sprecher, E., Trifonov, V., and Gerstein, M. (2007). The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. PLOS Computational Biology *3*, e59.

Yu, H., Tardivo, L., Tam, S., Weiner, E., Gebreab, F., Fan, C., Svrzikapa, N., Hirozane-Kishikawa, T., Rietman, E., Yang, X*., et al.* (2011). Next-generation sequencing to generate interactome datasets. Nat Meth *8*, 478-480.

Zhong, Q., Simonis, N., Li, Q.-R., Charloteaux, B., Heuze, F., Klitgord, N., Tam, S., Yu, H., Venkatesan, K., Mou, D*., et al.* (2009). Edgetic perturbation models of human inherited disorders. Mol Syst Biol *5*.

Zhu, Q., Ge, D., Maia, Jessica M., Zhu, M., Petrovski, S., Dickson, Samuel P., Heinzen, Erin L., Shianna, Kevin V., and Goldstein, David B. (2011). A Genome-wide Comparison of the Functional Properties of Rare and Common Genetic Variants in Humans. American Journal of Human Genetics *88*, 458-468.

CHAPTER 4

AN INTERACTOME PERTURBATION FRAMEWORK PRIORITIZES DAMAGING MISSENSE MUTATIONS FOR DEVELOPMENTAL DISORDERS

4.1 Preface

Chapter 4 has been accepted for publication as "Chen, S.*, **Fragoza, R.***, Klei, L., Liu, Y., Wang, J., Roeder, K., Devlin, B., Yu, H. An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders. *Nature Genetics*" where * indicates co-first author. Per guidelines by the Field of Biochemistry, Molecular, and Cellular Biology, The manuscript has been amended to focus primarily on contributions made by Robert Fragoza but has retained the results of others for the purposes of clarity when reading the text.

4.2 Abstract

Identifying disease-associated missense mutations remains a challenge, especially in large-scale sequencing studies. Here we establish an experimentally and computationally integrated approach to investigate the functional impact of missense mutations in the context of the human interactome network and test our approach by analyzing ~2,000 *de novo* missense (dnMis) mutations found in autism subjects and their unaffected siblings. Interaction-disrupting dnMis mutations are more common in autism probands, these mutations principally affect hub proteins, and they disrupt a significantly higher fraction of hub interactions than in unaffected siblings. Additionally, they tend to disrupt interactions involving genes previously implicated in

autism, providing complementary evidence that strengthens previously identified associations and enhances discovery of new ones. Importantly, by analyzing dnMis data from six disorders using the computational approach alone, we demonstrate that it offers a generalizable interactome-based framework for identifying and prioritizing missense mutations that contribute risk to human disease.

4.3 Introduction

Mutations disrupting the function of proteins are recognized as an important source of risk for developmental disorders (DDs), such as intellectual disability (Mefford et al., 2012; Ropers, 2010), autism spectrum disorder (ASD) (Devlin and Scherer, 2012) and congenital heart defects (Bruneau, 2008). Whole-exome sequencing (WES) has produced a boon of findings linking *de novo* mutations to risk for DDs (de Ligt et al., 2013; de Ligt et al., 2012; De Rubeis et al., 2014; Deciphering Developmental Disorders, 2015, 2017; Epi et al., 2013; Euro et al., 2014; Fromer et al., 2014; Gilissen et al., 2014; Iossifov et al., 2014; Iossifov et al., 2012; O'Roak et al., 2012; Rauch et al., 2012; Sanders et al., 2012; Zaidi et al., 2013). Not all mutations are simple to interpret as causing a loss of gene function. Missense mutations are especially difficult; although there are bioinformatics tools to predict the level of damage (Adzhubei et al., 2010; Kircher et al., 2014; Pollard et al., 2010), these annotators are far from perfect. This is a critical deficiency because the majority of coding mutations are missense. Here we show that one key feature in evaluating the disruptiveness of mutations is whether they fall in known or predicted protein-protein interaction interfaces and their likelihood to disrupt these interactions.

Large-scale studies of known disease-associated mutations have already reported a strong association with binding interfaces of protein interactions (Sahni et al., 2015; Wei et al., 2014). The major bottleneck for wide application of this feature is limited knowledge about the set of interactions and the binding interfaces of all interactions. To experimentally evaluate the impact of mutations on protein interactions, we establish a high-throughput mutagenesis and interactome-scanning pipeline for generating site-specific mutant clones and testing corresponding mutant protein interactions. Such a pipeline, however, cannot evaluate the impact of missense mutations on many interactions, because high-throughput interaction assays are limited in their coverage (Braun et al., 2009; Venkatesan et al., 2009; Yu et al., 2008). For this reason, we also explore a computational approach for systematically examining the functional impact of missense mutations on protein interactions. This approach builds on our newly-established full-interactome interface predictions (Meyer et al., 2018) to computationally predict the impact of all missense mutations on all associated interactions. Here we apply our experimental and computational approaches in tandem, which can be applied to any WES study.

To evaluate the effectiveness of our integrated experimental-computational approach, we focus on 2,821 *de novo* missense (dnMis) mutations identified from WES of ~2,500 families from the Simons Simplex Collection (SSC) (Sanders et al., 2015). The SSC targets the study of ASD through a cohort of parent-offspring trios or quads with two unaffected parents, an ASD proband and, for most families, an unaffected sibling (Fischbach and Lord, 2010). Previous analyses of the SSC data have reported significantly higher *de novo* mutation rate in ASD probands versus

unaffected siblings across various mutation types, from copy number variants (CNVs) (Levy et al., 2011; Sanders et al., 2011; Sanders et al., 2015), frameshift indels (Dong et al., 2014), to missense mutations (Iossifov et al., 2014; Iossifov et al., 2012; O'Roak et al., 2012; Sanders et al., 2012). While a number of risk *de novo* copy number (Levy et al., 2011; Pinto et al., 2010; Sanders et al., 2011; Sanders et al., 2015; Sebat et al., 2007) and protein truncating (Dong et al., 2014; Iossifov et al., 2014; Iossifov et al., 2012; O'Roak et al., 2012; Sanders et al., 2012) variants have been identified, exactly which dnMis mutations play a role and to what extent are open questions. We applied our integrated framework to evaluate the effect of 1,733 dnMis mutations within a protein interactome framework aiming to identify potentially disease-contributing dnMis mutations. Though there are many ways by which a missense mutation can impact a protein's function, such as by destabilizing protein folding, we evaluate the disruptiveness of a mutation within our framework exclusively on its capacity to disrupt protein interactions, measured experimentally or through prediction. We further compare the network properties of proteins impacted by interaction-disrupting and non-disrupting dnMis mutations, using unaffected siblings as negative controls throughout. While our analyses focus on dnMis mutations in ASD, the integrated experimental-computational approach provides a generalizable framework for investigating the impact of missense mutations uncovered by WES for human diseases.

4.4 Results

4.4.1 Proband dnMis mutations are enriched on interaction interfaces

We previously reported that inherited in-frame disease-associated mutations are significantly enriched on protein interaction interfaces and demonstrated that alteration of specific protein interactions is crucial in the pathogenesis of many disease-associated genes (Wang et al., 2012). To explore the relationship between non-inherited dnMis mutations and autism, we used a structurally-resolved 3D human interactome network (Meyer et al., 2013; Wang et al., 2012) to examine where dnMis mutations reside with respect to interaction interfaces. We found that in probands, dnMis mutations are significantly enriched on interaction interfaces: while interaction interfaces cover 30.1% of the proteins harboring these mutations, 38.2% of the mutations fall in interaction interfaces (1.27 fold, $P = 2.9\times10^{-3}$ by two-tail exact binomial test). In contrast, dnMis mutations in siblings fall in interaction interfaces on corresponding proteins at an expected rate (observed 37.6% versus expected 36.5%, 1.03 fold, $P = 0.76$). Thus, disruption of specific interactions could contribute to ASD etiology for dnMis mutations in probands (**Figure 4.4-1A**), underscoring the functional significance of dnMis mutations on protein interaction interfaces.

4.4.2 Proband dnMis mutations are more disruptive than sibling mutations

We next explored the impact of dnMis mutations on protein interactions by intersecting all 2,821 dnMis mutations with 59,073 human protein interactions from a comprehensive set of high-quality physical interactions compiled from eight widely-used interaction databases (Das and Yu, 2012), including BioGRID (Chatr-Aryamontri

et al., 2015), MINT (Stelzl et al., 2005), iRefWeb (Turner et al., 2010), DIP (Salwinski et al., 2004), IntAct (Hermjakob et al., 2004), HPRD (Keshava Prasad et al., 2009), MIPS (Mewes et al., 2011), and the PDB (Berman et al., 2000). Of these mutations, 1,733 are on proteins with at least one known interaction within the current human interactome dataset. To experimentally assess the impact of a subset of these mutations, 208 individual clones were generated carrying dnMis mutations – corresponding to 109 in probands and 99 in siblings, respectively – using Clone-seq, a massively parallel site-directed mutagenesis pipeline (Wei et al., 2014) (**Materials and Methods**). Protein interactions amenable to yeast two-hybrid (Y2H) were then tested, yielding 667 total protein interactions corresponding to 151 of our cloned dnMis mutations (**Figure 4.4-1B; Materials and Methods**).

To explore the remaining dnMis mutations and interactions untested by Y2H, we applied a two-tiered computational approach that first predicts whether a particular residue is an interface residue using Interactome INSIDER (Meyer et al., 2018), a unified machine-learning framework comprising the first full-interactome map of human interaction interfaces. To determine whether a particular mutation is deleterious, we used PolyPhen-2 (PPH2) (Adzhubei et al., 2010) predictions: if a particular residue is predicted to be an interface residue and its mutation is scored as "probably damaging" by PPH2, that mutation was predicted as *interaction-disrupting*; if a mutation is unlikely to occur at an interface residue and is scored as "benign" by PPH2, it was predicted as *interaction non-disrupting* (**Figure 4.4-1B; Materials and Methods**).

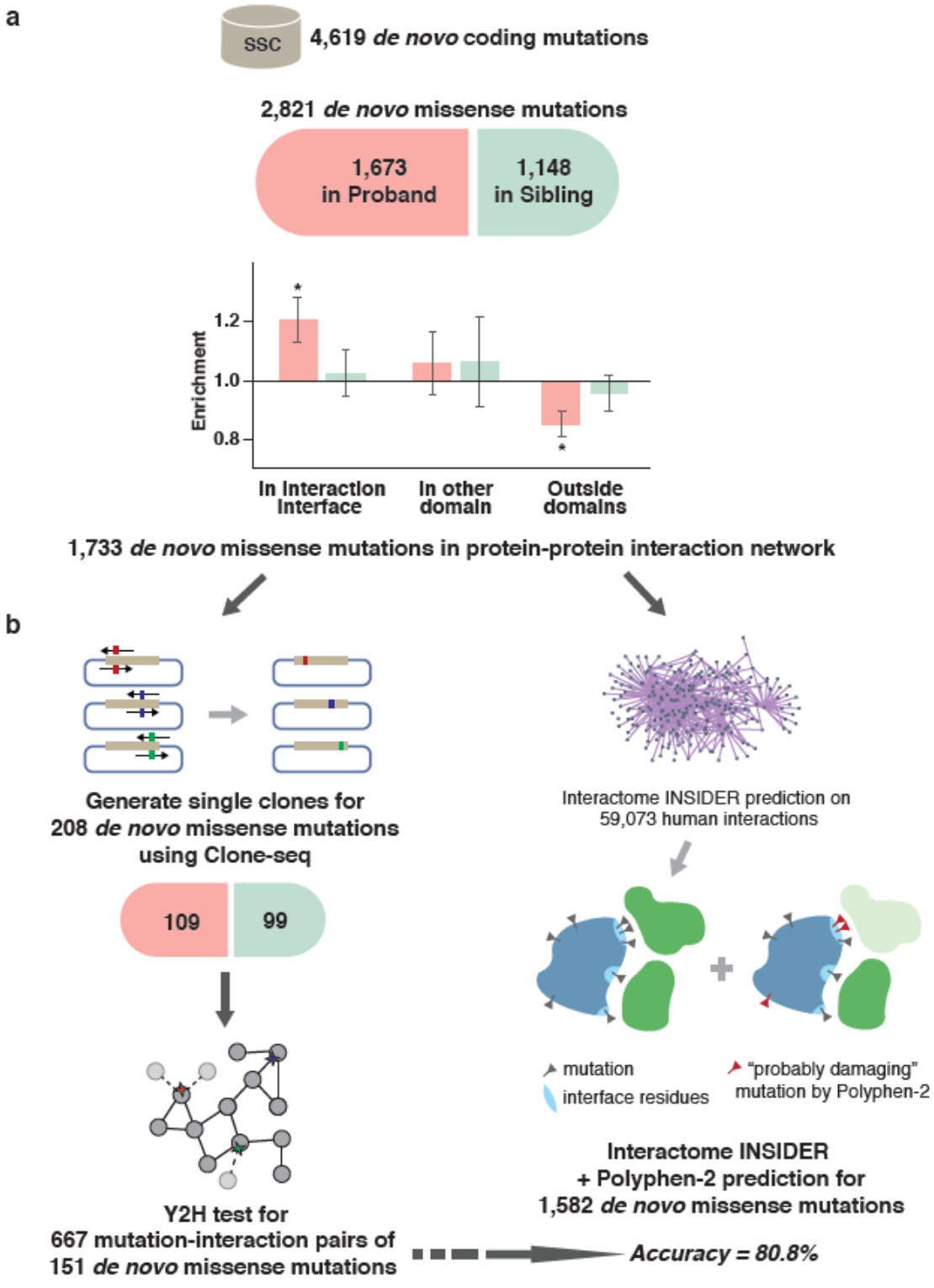


Figure 4.4-1 Workflow of our integrated experimental-computational interactome perturbation framework. *(a) Enrichment distribution of dnMis mutations from SSC in*

*different locations of interacting proteins. Enrichment was calculated by the ratio of the observed fraction of dnMis mutations that occur on interaction interfaces over the fraction of interface residues on corresponding proteins (expected fraction). P-values calculated using two-tail exact binomial test (*P < 0.05, Online Methods). Error bars indicate ± standard error. (b) Experimental (left) and computational (right) pipelines for assessing the functional impact of dnMis mutations on protein-protein interactions.*

To evaluate the performance of our computational prediction, we applied this two-tiered prediction approach to our 667 experimentally tested protein interactions and obtained an accuracy of 80.8% (sensitivity: 65.0%, specificity: 82.5%). Additionally, when our approach was applied to a previously published, independent dataset of 204 disease-associated mutations and their impact on protein-protein interactions (Wei et al., 2014), we obtained a similar prediction performance (accuracy: 77.4%, sensitivity: 81.0%, specificity: 75.0%).

We then analyzed the distribution of disrupted interactions across ASD probands and unaffected siblings. Examining our experimental data revealed that 74/361 (20.5%) tested interactions were disrupted in probands. In contrast, only 21 out of 208 (10.1%) interactions were disrupted in unaffected siblings. Modeling the count of disruptions per subject with a negative binomial model, using case status as the predictor, yielded a 2.54-fold higher rate of disruptions in probands ($P = 0.012$, **Figure 4.4-2A; Materials and Methods**). This sharp contrast in interaction disruption rate suggests that disruption of the interactome network by dnMis mutations contributes to autism etiology in probands. Combining the experimental data with predictions for all remaining dnMis mutations and interactions, there was again a significant, 2.34-fold higher disruption rate for probands (22.8%) versus unaffected

siblings (8.7%, $P = 5.6\times10^{-4}$, **Figure 4.4-2B**). Furthermore, the predicted disruptions alone showed significantly higher rate of disruption in probands than siblings (2.15 fold, $P = 0.013$, **Figure 4.4-2C**).

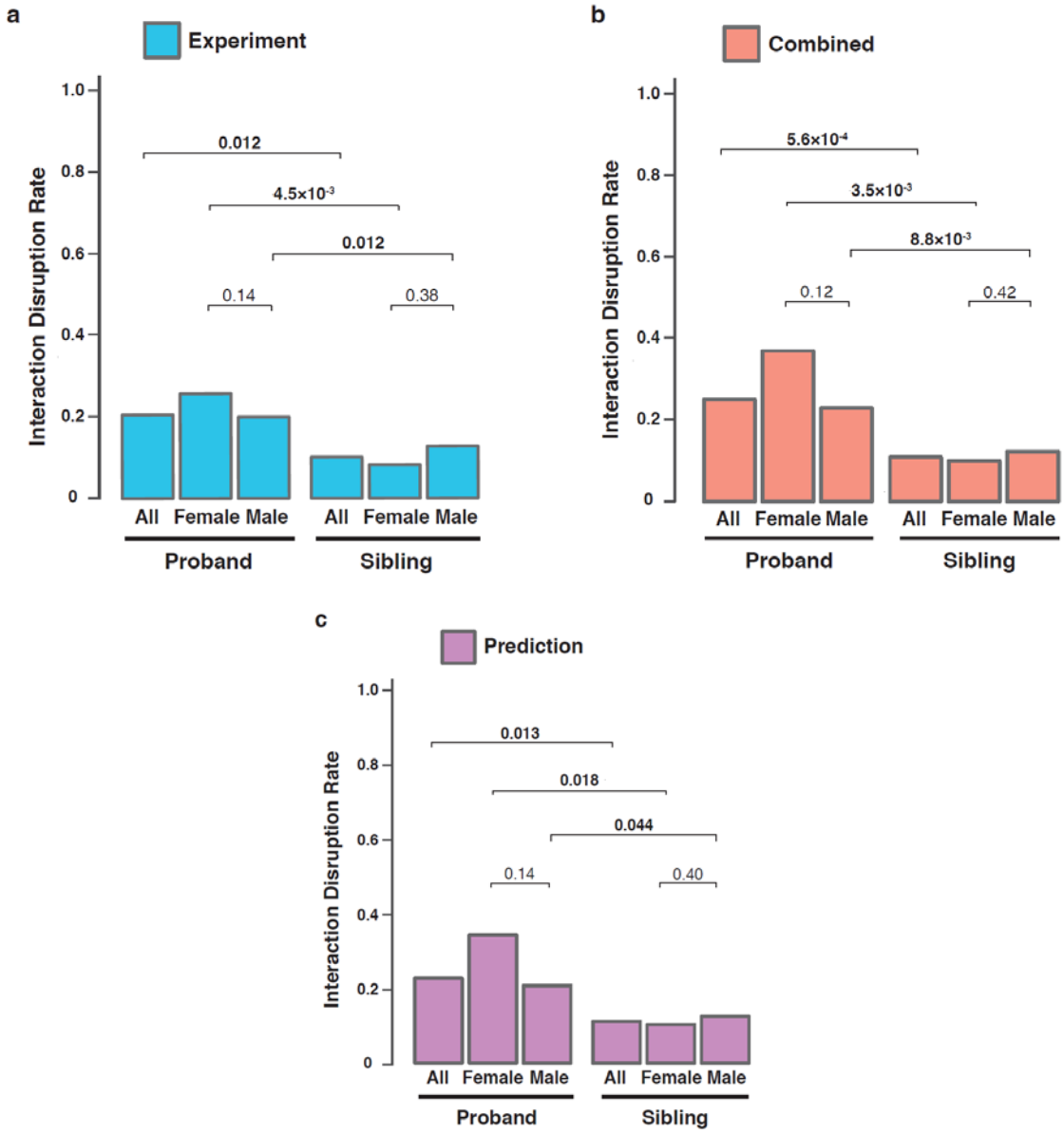


Figure 4.4-2 dnMis mutations are more disruptive in ASD probands than in siblings. *Interaction disruption rates of dnMis mutations (a) tested experimentally, (b) by combining experimental results and predictions, and (c) predicted computationally. Probands and*

*unaffected siblings are divided by sex. The count of disruptions per subject was modeled with a negative binomial model (P < 0.05 in **bold**). Combined: 1,080 out of 4,275 measured interactions were disrupted in ASD probands (25.3%) and 322 out of 2,973 were disrupted in unaffected siblings (10.8%). The interaction disruption rate is significantly higher in ASD probands than that in unaffected siblings ($P = 5.6\times10^{-4}$, FC = 2.34 [1.44–3.79, 95% CI], two-tail negative binomial model). The trend persists in male and female subgroups: 23.1% disruption rate in male ASD probands versus 12.3% in male siblings ($P = 8.8\times10^{-3}$, FC = 2.21 [1.15–4.25, 95% CI], one-tail negative binomial model); 37.3% disruption rate in female ASD probands versus 9.9% in female siblings ($P = 3.5\times10^{-3}$, FC = 3.50 [1.41–8.72, 95% CI]). Comparing disruption rates between males and females revealed a higher rate, although not quite significant, in females than males in ASD probands (P = 0.12, FC = 1.71 [0.71–4.09, 95% CI], one-tail negative binomial model), whereas similar rates were observed in female and male siblings (P = 0.42, FC = 1.08 [0.52–2.22, 95% CI]).*

These observations suggest that dnMis mutations in ASD probands are of higher functional consequence than those in unaffected siblings. Therefore, interaction-disrupting mutations identified by our integrated experimental-computational framework could serve as a viable approach for identifying candidate risk variants, which may go undetected by other methodologies. Hereinafter, we shall present results using the combined data. Results using only the Y2H data or predictions are provided in **Appendix D.1**.

The female protective effect postulates that females require a larger genetic burden before being diagnosed with ASD (Chang et al., 2015; De Rubeis et al., 2014). Accordingly, we anticipate dnMis mutations in female probands to be more disruptive than those in male probands, although the 6.5:1 male:female ratio of probands could obscure true differences by limiting power. Indeed, we observed a higher disruption rate in females than in males among ASD probands, fold = 1.71, but the difference is not quite significant ($P = 0.12$). In contrast, the disruption rate in female versus male

siblings is 1.08-fold and does not approach significance (P = 0.42, **Figure 4.4-2B**).

4.4.3 Disruptive dnMis mutations in probands principally impact network hubs

Previous research has shown that genes harboring known disease-associated mutations differ strongly in their network properties in comparison to non-disease-associated genes (Feldman et al., 2008; Goh et al., 2007). Early studies reported that disease-associated genes often encode for protein hubs that mediate a greater number of protein interactions than their non-disease associated counterparts as a whole (Jonsson and Bates, 2006; Xu and Li, 2006). However, researchers later argued that the observed hub-disease gene correlation might be entirely driven by a handful of hub-encoding essential genes classified within the disease-associated gene class (Goh et al., 2007). Here we investigated whether proteins harboring disruptive dnMis mutations in ASD probands exhibit distinguishable network properties in the human interactome.

We first compared the degree of all proteins harboring interaction-disrupting dnMis mutations to those harboring non-disrupting dnMis mutations. We found that in ASD probands, proteins with interaction-disrupting dnMis mutations on average have a significantly higher degree than proteins with non-disrupting dnMis mutations (mean$\pm$s.e.m: 18.4$\pm$2.8 versus 9.3$\pm$1.0, fold change [FC] = 1.98, $P = 2.0\times10^{-6}$ by two-tail U-test, **Figure 4.4-3A**), whereas no significant difference was observed in unaffected siblings (mean$\pm$s.e.m: 7.9$\pm$1.0 versus 11.4$\pm$1.3, FC = 0.69, P = 0.60, **Figure 4.4-3A**).

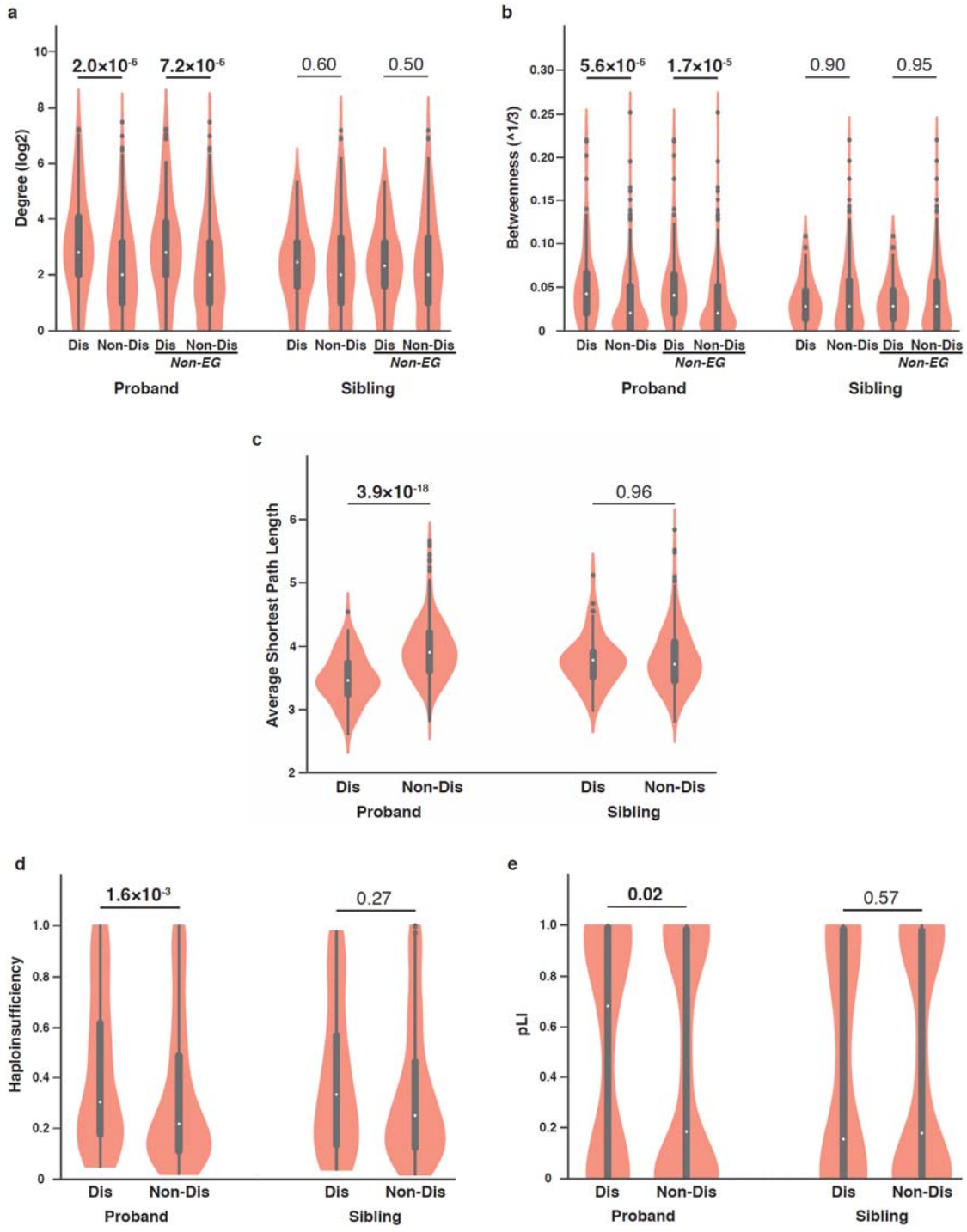


Figure 4.4-3 Disruptive proband dnMis mutations exhibit characteristic network and haploinsufficiency properties. *(a) Degree and (b) betweenness distributions of proteins with interaction-disrupting (Dis, n = 109 in probands and n = 68 in siblings) or non-disrupting (Non-Dis, n = 342 in probands and n = 241 in siblings) dnMis mutations across all proteins*

and across non-essential gene-encoded proteins (Non-EG) in ASD probands (Dis: n = 106; Non-Dis: n = 338) and unaffected siblings (Dis: n = 66; Non-Dis: n = 238). (c) Average shortest path length distributions of proteins with dnMis mutations (in probands, Dis: n = 109, Non-Dis: n = 342; in siblings, Dis: n = 68, Non-Dis: n = 241). (d) Haploinsufficiency and (e) pLI distributions of genes with dnMis mutations. Genes with available haploinsufficiency or pLI scores were included in corresponding analyses (haploinsufficiency in probands, Dis: n = 95, Non-Dis: 304; in siblings, Dis: n = 63, Non-Dis: n = 217; pLI: in probands, Dis: n = 106, Non-Dis: n = 338; in siblings, Dis: n = 63, Non-Dis: n = 237). Proteins with dnPTVs were excluded from all above analyses. Violin plots: thick black bar, interquartile range; white dot, median; whiskers, upper and lower limits; points, outliers; while the width of each 'violin' is proportional to element abundance. P-values were calculated using two-tail U-test (P < 0.05 in ***bold****).*

This suggests that interaction-disrupting dnMis mutations in ASD probands preferentially impact hub proteins, which play a central role in maintaining the integrity of the human interactome (Albert et al., 2000).

Importantly, when we excluded essential human genes (Chen et al., 2017) from our analysis, the correlation between interaction-disrupting dnMis mutations and protein hubs persisted (mean±s.e.m: 17.6±2.9 versus 9.2±1.0, FC = 1.91, $P = 7.2\times10^{-6}$, **Figure 4.4-3A**). Similarly, no such correlation in unaffected siblings was observed (mean±s.e.m: 8.0±1.0 versus 11.0±1.3, FC = 0.73, $P = 0.50$, **Figure 4.4-3A**). Likewise, when we analyzed betweenness, another measure of network centrality based on shortest paths, proteins harboring interaction-disrupting dnMis mutations have a significantly higher betweenness value than proteins harboring non-disrupting dnMis mutations, regardless of whether essential genes were included (**Figure 4.4-3B**).

To further assess whether disruptive dnMis mutations tend to be on essential

genes, we analyzed gene essentiality measured by Wang *et al.* using CRISPR gene knockout screens (Wang et al., 2015). Using this CRISPR score, we observed no significant difference in essentiality between genes with interaction-disrupting and non-disrupting dnMis mutations for probands (mean±s.e.m: -0.43 ±0.08 versus -0.33±0.04, FC = 1.30, $P = 0.28$ by two-tail *U*-test) or for unaffected siblings (mean±s.e.m: = -0.37±0.09 and -0.41±0.05, FC = 0.90, respectively; $P = 0.39$). This confirms that disruptive dnMis mutations have no tendency to be on essential genes while preferentially affecting topologically central positions in the interactome network.

We then investigated whether proteins with dnMis mutations tend to form inter-connected modules within the interactome network. We found that in ASD probands, proteins with interaction-disrupting dnMis mutations on average have significantly smaller shortest path length to each other than that of proteins harboring non-disrupting dnMis mutations (mean±s.e.m: 3.48±0.04 versus 3.94±0.03, FC = 0.88, $P = 3.9\times10^{-18}$ by two-tail *U*-test, **Figure 4.4-3C**). This result indicates that proteins with disruptive dnMis mutations in probands tend to be closely connected to each other in the network and may therefore function as modules with specific roles in ASD etiology. In contrast, no such trend was observed for proteins with disruptive dnMis mutations in unaffected siblings (mean±s.e.m: 3.77±0.05 versus 3.79±0.03, FC = 0.99, $P = 0.96$, **Figure 4.4-3C**), underscoring the functional significance of modules derived from interaction-disrupting dnMis mutations in ASD probands.

Overall, our analyses indicate that network topology should be considered when interpreting the impact of dnMis mutations. In this manner, we can investigate

how disruptive missense mutations alter local community structure and how information flow through multiple mutations could work together to rewire the whole network that can lead to autism or other disease-associated phenotypes.

4.4.4 Disruptive dnMis mutations in probands target haploinsufficient genes

Disruptive dnMis mutations typically occur only on one copy of the gene. To affect risk, they should occur more frequently on haploinsufficient genes, where a single copy of the wild-type gene is insufficient to carry out its normal function. In probands, genes harboring interaction-disrupting dnMis mutations had a higher probability of being haploinsufficient (Huang et al., 2010) than genes harboring non-disrupting dnMis mutations (mean±s.e.m: 0.42±0.03 versus 0.33±0.02, FC = 1.27, $P = 1.6\times10^{-3}$ by two-tail *U*-test, **Figure 4.4-3D**). In contrast, no significant difference was observed in unaffected siblings (mean±s.e.m: 0.39±0.04 versus 0.34±0.02, FC = 1.15, P = 0.27, **Figure 4.4-3D**). Reinforcing these findings, we also found that genes with interaction-disrupting dnMis mutations in probands are, as a whole, less tolerant to genetic variation, as indicated by their higher average pLI(Lek et al., 2016) scores in comparison to genes with non-disrupting dnMis mutations (mean±s.e.m: 0.52±0.04 versus 0.43±0.02, FC = 1.21, P = 0.02, **Figure 4.4-3E**). No such contrast was found in unaffected siblings (mean±s.e.m: 0.44±0.06 versus 0.44±0.03, FC = 1.00, P = 0.57, **Figure 4.4-3E**). Collectively, these results demonstrate that interaction-disrupting dnMis mutations in ASD probands tend to affect haploinsufficient genes, for which heterozygous variations are not tolerated, and they may therefore contribute to ASD outcomes through dosage effect (Ronemus et al., 2014).

4.4.5 Disruptive dnMis mutations in probands cluster closely to known ASD genes

To evaluate whether interaction-disrupting dnMis mutations are associated with ASD risk, we first investigated whether such mutations are enriched in previously reported ASD-associated genes. Using a curated list of 881 genes implicated in ASD in the SFARI database (Basu et al., 2009), we observed a significant enrichment in probands for genes with interaction-disrupting dnMis mutations compared to genes with non-disrupting dnMis mutations (21/109 versus 32/342, OR = 2.3, $P = 5.7\times10^{-3}$ by one-tail Fisher's exact test). In contrast, no significant overlaps with ASD-associated genes for interaction-disrupting dnMis mutations were observed in unaffected siblings (6/68 versus 17/241, OR = 1.3, $P = 0.39$). Thus, characterizing interaction perturbation captures new evidence to establish associations of genes with ASD.

Previous studies have reported functional clustering in genes with dnPTVs in ASD individuals (Iossifov et al., 2012; Neale et al., 2012a; O'Roak et al., 2012; Sanders et al., 2012). Here we assessed the network distance within the human interactome between genes harboring interaction-disrupting dnMis mutations (excluding genes with dnPTVs) and seven classes of known ASD-associated genes. These genes include: (1) FMRP target genes, with transcripts bound by the fragile X mental retardation protein (FMRP); (2) genes encoding chromatin modifiers (CHM); (3) genes expressed preferentially in embryos (EMB); (4) genes encoding postsynaptic density proteins (PSD); (5) 881 genes in the SFARI database; (6) a high-quality SFARI subset (SFARI-hq, 141 genes scored as syndromic, high confidence, or strong candidate (Basu et al., 2009)); and (7) the latest set of 65 ASD genes discovered by *de*

novo mutations (DN65) (Sanders et al., 2015). We found that in probands, proteins harboring interaction-disrupting dnMis mutations are significantly closer to proteins from all seven classes in comparison to proteins with non-disrupting dnMis mutations (**Table 4.4-1; Materials and Methods**). In contrast, no significant differences were observed among unaffected siblings in any category. These findings demonstrate that disruptive dnMis mutations identified by our study are indeed closely related to known ASD genes and functional classes and that they may contribute to ASD etiology by disrupting common pathways shared with dnPTVs.

	Proband					**Sibling**				
	Dis (101)		**Non-Dis (319)**		***P*-value**	**Dis (63)**		**Non-Dis (220)**		***P*-value**
	Mean	**s.e.m**	**Mean**	**s.e.m**		**Mean**	**s.e.m**	**Mean**	**s.e.m**	
FMRP (794)	2.61	0.04	2.85	0.03	***1.5e-6***	2.76	0.05	2.77	0.03	*0.57*
CHM (408)	2.55	0.04	2.79	0.03	***1.3e-6***	2.70	0.05	2.72	0.03	*0.44*
EMB (1,865)	2.65	0.04	2.88	0.03	***2.9e-6***	2.79	0.05	2.81	0.03	*0.45*
PSD (1,395)	2.61	0.04	2.84	0.03	***2.4e-6***	2.76	0.05	2.77	0.03	*0.47*
SFARI (881)	2.69	0.04	2.92	0.03	***1.8e-6***	2.83	0.05	2.85	0.03	*0.52*
SFARI hq (141)	2.62	0.04	2.86	0.03	***1.1e-6***	2.77	0.05	2.77	0.03	*0.58*
DN65 (65)	2.70	0.04	2.94	0.03	***1.0e-6***	2.85	0.05	2.86	0.03	*0.52*

Table 4.4-1 Distance of proteins with interaction-disrupting (Dis) and non-disrupting (Non-Dis) dnMis mutations to seven classes in a protein interactome network background.
Number of proteins in each class is indicated in parentheses. Proteins with dnPTVs in probands and siblings were excluded from the analyses. P-values were calculated using one-tail U-test (P < 0.05 in ***bold****).*

4.4.6 Identification of candidate ASD genes and mutations

Towards the identification of new candidate ASD-associated genes, we examined mutations on the protein RARA. RARA binds with RXRB to form the retinoic acid (RA) receptor complex. When bound to RA, the retinoic acid receptor can then bind RA receptor elements (RAREs) to co-activate transcription of downstream genes. In agreement with our Y2H experiments, our computational approach predicted that a proband mutation p.Pro375Leu on RARA is disruptive, while an unaffected sibling mutation, p.Arg83His, is not (**Figure 4.4-4A** and **Figure 4.4-4B**). We note that PPH2 predicts both mutations to be probably damaging and cannot distinguish the two. We further confirmed by co-immunoprecipitation in human cells that the proband mutation p.Pro375Leu disrupts the RARA-RXRB interaction while the sibling mutation p.Arg83His does not (**Figure 4.4-4C; Materials and Methods**).

While there is insufficient evidence to directly link mutations on RARA to ASD, there is compelling evidence that mutated RARA does induce ASD risk by affecting RA signaling. Specifically, we would expect the p.Pro375Leu mutation to diminish RA signaling by disrupting its binding to RXRB. Notably, one of the most common genetic risk factors for ASD is maternal duplication of 15q11-q13 and isodicentric chromosome 15 (Schanen, 2006), both of which increase transcription of *UBE3A*, among other genes. It has recently been shown that UBE3A negatively regulates ALDH1A proteins (Xu et al., 2017), which is the rate-limiting enzyme of RA synthesis. Increased dosage of UBE3A diminishes RA synthesis and RA signaling, altering neuronal development and features such as homeostatic synaptic plasticity (Xu et al., 2017).

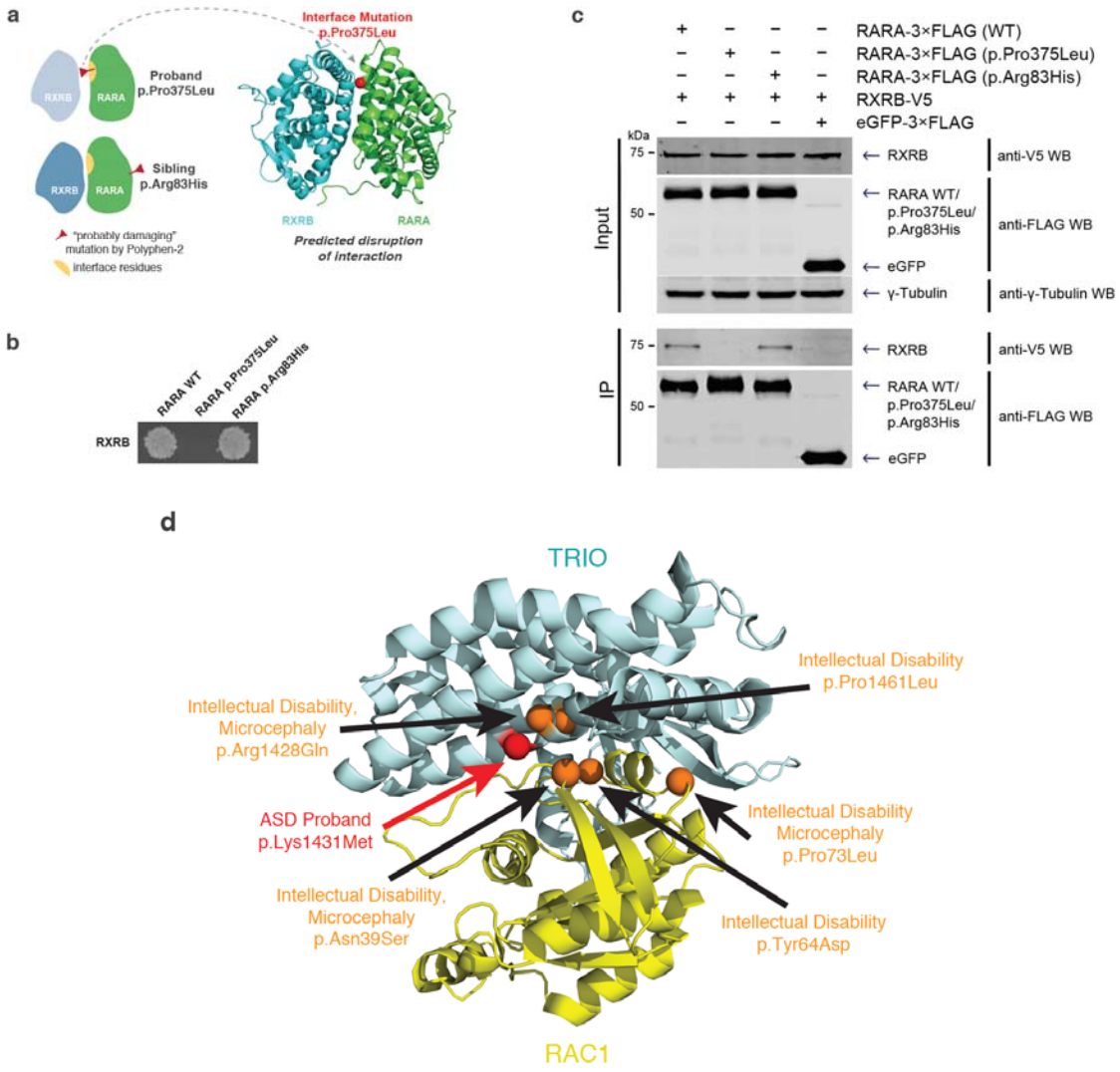


Figure 4.4-4 Identification of candidate ASD-associated genes and mutations through our interactome perturbation framework. *(a) Computational prediction of the effects of RARA p.Pro375Leu and RARA p.Arg83His on the RARA-RXRB interaction. A homology model highlighting the RARA p.Pro375Leu interface mutation is shown. (b) RARA p.Pro375Leu disruption and RARA p.Arg83His non-disruption of RARA-RXRB interaction by Y2H. (c) Co-immunoprecipitation confirming RARA p.Pro375Leu disruption and RARA p.Arg83His non-disruption of RARA-RXRB interaction in HEK 293T cells. See Supplementary Fig. 10 for uncropped gel images. (d) Co-crystal structure of TRIO-RAC1 displaying the structural locations of proband ASD (red) and intellectual disability and/or microcephaly (orange) dnMis mutations across the interaction interface.*

Moreover, in mice, ASD-like phenotypes are induced by over-expression of UBE3A or by an ALDH1A antagonist, while the wild-type phenotype can be rescued by RA supplementation (Xu et al., 2017). Thus, together with published results regarding the role of UBE3A in RA signaling and autism risk (Xu et al., 2017), our results implicate RARA as an ASD-associated gene, and our experimentally-validated interaction-disrupting prediction for RARA p.Pro375Leu demonstrates how our methodology can be used to identify functional dnMis mutations.

The occurrence of a predicted disruptive mutation near other closely related disease-associated dnMis mutations across interacting proteins can lend strong evidence towards the postulated functionality and shared phenotypic impact of the mutation in question. In this regard, our computational approach predicted an ASD proband mutation, p.Lys1431Met, on the guanine nucleotide exchange factor TRIO that disrupts its interaction with the GTPase RAC1 (**Figure 4.4-4D**). Of note, two neurodevelopmental disorder dnMis mutations (Deciphering Developmental Disorders, 2017; Pengelly et al., 2016) on TRIO, p.Arg1428Gln and p.Pro1461Leu, occur in structural proximity to the ASD proband interface mutation, p.Lys1431Met, as do three dnMis mutations on RAC1 interface residues, p.Asn39Ser, p.Tyr64Asp, and p.Pro73Leu, which all result in mild to severe intellectual disability (Reijnders et al., 2017) (**Figure 4.4-4D**). Moreover, p.Lys1431Met has been recently reported to functionally inhibit synaptic function in human cell lines and statistically postulated to reside within a hotspot for ASD-related *de novo* mutations in the GEF1 domain of TRIO (Sadybekov et al., 2017). As sequencing data from DD studies becomes more readily available, we anticipate the use of predicted interaction-disrupting mutations to

uncover shared molecular pathways between related DDs.

4.4.7 An excess of dnMis mutations in DDs occur on interaction interfaces

To demonstrate the generalizability of our interactome perturbation approach towards studying the impact of missense mutations in human disease, we investigated how ~10,000 dnMis mutations previously detected in DDs correspond with protein interaction interfaces. The mutation data comprises a collection of 4,565 dnMis mutations from the Deciphering Developmental Disorders project and five lists of dnMis mutations curated from studies of autism, congenital heart disease, intellectual disability, schizophrenia, and epilepsy (denovo-db v.1.5) (Turner et al., 2017). We found that in all six datasets, dnMis mutations occur significantly more frequently on protein interaction interfaces than expected (**Figure 4.4-5**), indicating that dnMis mutations in DDs can contribute to disease risk by impacting protein interactions. In particular, the strongest signal was observed in intellectual disability: 23.5% of the dnMis mutations occurred on interaction interfaces, resulting in an enrichment of 2.09 (1.77–2.44, 95% CI) in comparison to the fraction of interface residues on corresponding proteins (11.2%, $P = 1.4\times10^{-14}$ by two-tail exact binomial test). In contrast, dnMis mutations in schizophrenia had the weakest significance ($P = 8.7\times10^{-4}$, Enrichment = 1.61 [1.22-2.06, 95% CI]), which agrees with previous findings that schizophrenia has a much weaker *de novo* signal than other DDs (Purcell et al., 2014).

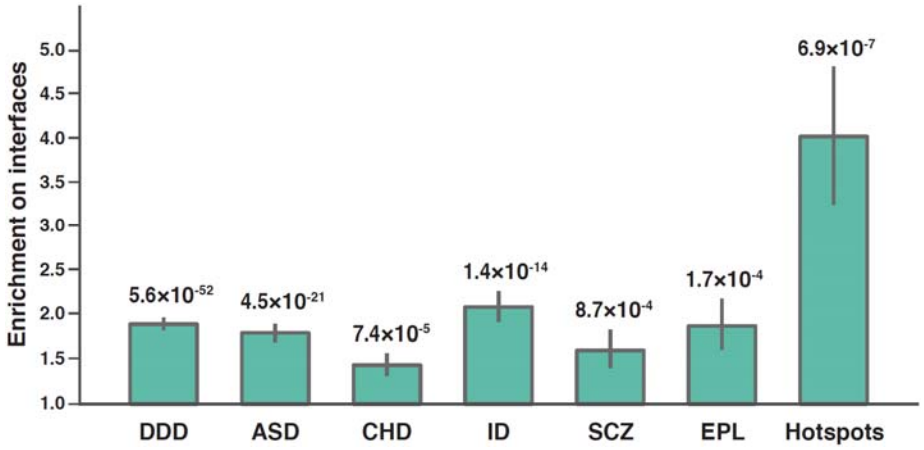


Figure 4.4-5 dnMis mutations are enriched on protein interaction interfaces in developmental disorders. *Enrichment was calculated by the ratio of the observed fraction of dnMis mutations that occur on interaction interfaces over the fraction of interface residues on corresponding proteins (expected fraction). Error bars indicate ± standard error. P-values were calculated using two-tail exact binomial test. DDD (Deciphering Developmental Disorders project, n = 2,914 dnMis mutations): Enrichment = 1.90 (1.76-2.04, 95% CI); ASD (autism spectrum disorder, n = 1,512): Enrichment = 1.80 (1.61–2.00, 95% CI); CHD (congenital heart disease, n = 759): Enrichment = 1.44 (1.21–1.70, 95% CI); ID (intellectual disability, n = 498): Enrichment = 2.09 (1.77–2.44, 95% CI); SCZ (schizophrenia, n = 312): Enrichment = 1.61 (1.22-2.06, 95% CI); EPL (epilepsy, n = 181): Enrichment = 1.88 (1.36–2.48, 95% CI); Hotspots (n = 31): Enrichment = 4.03 (2.51–5.58, 95% CI).*

Geisheker *et al*. recently reported 40 dnMis hotspots implicated in neurodevelopmental disorder pathogenesis (Geisheker et al., 2017). When we examined the 31 corresponding hotspots within the interactome network, we found that they occur on protein interaction interfaces at a very high rate of 48.4% (Enrichment = 4.03 [2.51–5.58, 95% CI], $P = 6.9\times10^{-7}$, **Figure 4.4-5**). This suggests that interactome perturbations play an important role in the pathogenesis linked with these recurrent events. Taken together, these findings reinforce that our integrated experimental-computational interactome perturbation approach offers a scalable and

generalizable framework to identify risk dnMis mutations in human disease.

4.5 Discussion

Here we demonstrated that dnMis mutations can contribute to ASD risk by disrupting protein-protein interactions and that our interactome perturbation framework offers a novel and effective way to identify ASD risk dnMis mutations. Because only a small fraction of dnMis mutations found in ASD subjects are believed to be functional (Iossifov et al., 2014), this framework helps overcome a significant challenge in identifying risk dnMis mutations. Our analyses focused on dnMis mutations from the SSC families because the information on unaffected siblings in the dataset provides robust negative controls. Our results demonstrated that interaction-disrupting dnMis mutations in ASD probands preferentially impact proteins that have many interaction partners in the interactome network (i.e., hubs) and disrupt these interactions at a significantly higher rate than those in unaffected siblings. Our results also lend evidence to previously reported ASD-associated genes and pathways by showing that interaction-disrupting dnMis mutations are closely clustered to proteins in ASD-associated functional classes in the interactome network. Thus, characterizing interactome perturbation provides additional and potentially orthogonal information to strengthen previously identified genetic associations and helps discover new genes that contribute to ASD risk.

Integration of computational predictions with experimental data imbued far more meaning onto missense mutations found in ASD probands and their siblings. Thus the prediction model alone can enhance researchers' ability to prioritize

damaging missense mutations and can be applied across a wide range of human disease studies. We emphasize that the strength of this prediction model is rooted in its integration of PPH2 scores and Interactome INSIDER interface predictions. To demonstrate this, we repeated all analyses using PPH2 and Interactome INSIDER separately and found that neither method individually is sufficient to reproduce most signals towards identifying disease-contributing dnMis mutations in ASD. This confirms that our two-tiered predictor, which evaluates the disruptiveness of a variant on protein interactions, greatly improves the effectiveness of predicting functional missense mutations. Taken together, we demonstrate that our computational prediction approach can serve as an effective and robust method to identify disease-contributing missense mutations.

Our analyses indicate that network properties are important in interpreting the functional impact of dnMis mutations and their relevance towards disease etiology. However, we recognize that the human interactome with which these analyses are performed is currently incomplete. As a result, certain classes of protein interactions, for example interactions mediated by membrane-bound proteins, may be under-represented in the current interactome, limiting potential insights from such proteins. Moreover, literature-derived segments of the human interactome are subject to sampling bias present in small-scale studies (Das and Yu, 2012; Rolland et al., 2014). Therefore, we re-examined the network topology analyses across a chronologically-ordered series of unbiased high-throughput (HT)-derived human interactomes. We show that not only are our results robust across all HT-derived interactomes, more importantly, we also found that the topological difference between interaction-

disrupting and non-disrupting dnMis mutations in probands becomes more significant as the interactome coverage increases. Taken together, we fully expect that as increasingly more human protein-protein interactions and mutations are uncovered, our interactome-perturbation framework can be applied to these new interactions and mutations to identify new or currently under-characterized disease-associated mutations and genes.

As large-scale WES studies continue to produce mutation data at ever-increasing scales, our interaction-disruption prediction approach can greatly extend the reach of interactome perturbation studies for investigating complex genotype-phenotype relationships and improving our understanding of how genetic variation affects disease risk through the alteration of topological and community structures of networks.

4.6 Materials and Methods

4.6.1 Enrichment of dnMis mutations on interaction interfaces

The set of 412 proteins with dnMis mutations and containing at least one interaction interface and one known domain was included for calculating dnMis mutation distribution. The sequences were divided into three regions: “in interaction interface”, “in other domain” and “outside domains”. Interaction interfaces were determined by our previously developed human structural interaction network [hSIN (Wang et al., 2012), comprising 4,222 structurally resolved interactions between 2,816 proteins]. Other domains were referred to protein domains [obtained from Pfam (Finn et al., 2016) database] that exclude interacting interfaces in hSIN. The rest of residues then

were categorized as “outside domains”. If the locations of mutations were not influenced by the domain architecture of the protein, then their relative lengths should determine the frequency of mutations in these three regions. The fraction of mutations expected by chance in each region was calculated by adding the total sequence length of each region in all proteins, and dividing it by the length of all proteins combined; call the probability of falling in an interaction interface, p. The number of observed mutations in each region over all proteins was also computed, call the number falling in the interaction interfaces S, and let N be the total number of missense mutations. An exact binomial test was then computed from p, S, and N. Confidence intervals (CIs) are based on 95% CI for an exact binomial, then transformed to the risk ratio (Enrichment) using the expectation in the denominator and the lower/upper bound in the numerator.

In ASD probands, the total length of 248 proteins is 377,421, which compromises of 113,449 residues on interaction interface, 69,870 residues on other domains, and 194,102 residues outside domains. The probabilities for a mutation to fall in these regions then were computed to be 30.1%, 18.5%, and 51.4%, respectively. The observed distribution of the 296 dnMis mutations on these proteins was 113 on interface, 59 on other domains, and 124 outside domains, revealing that dnMis mutations in ASD probands are significantly enriched on protein interaction interface ($P = 2.9\times10^{-3}$, Enrichment = 1.27 [1.09–1.46, 95% CI]) while occur on other domains with expected rate ($P = 0.55$, Enrichment = 1.08 [0.84–1.35, 95% CI]) and are depleted from regions outside domains ($P = 1.1\times10^{-3}$, Enrichment = 0.81 [0.70–0.93, 95% CI]). In contrast, the observed 186 dnMis mutations in unaffected siblings occur

on all three regions with expected rates: 70/186 fall on interface (37.6% versus expected rate of 36.5%, $P = 0.76$, Enrichment = 1.03 [0.84–1.23, 95% CI]), 32/186 fall on other domains (17.2% versus 15.9%, $P = 0.62$, Enrichment = 1.08 [0.76–1.47, 95% CI]), and 84/186 fall outside domains (45.2% versus 47.6%, $P = 0.51$, Enrichment = 0.95 [0.79–1.10, 95% CI]).

4.6.2 Cloning of 208 dnMis mutations using our massively-parallel Clone-seq pipeline

Single colony-derived mutant clones were constructed using a high-throughput mutagenesis and next-generation sequencing pipeline called Clone-seq (Wei et al., 2014). Wild-type clones were picked from hORFeome v8.1 (Yang et al., 2011) to serve as templates for site-directed mutagenesis (Eurofins). Mutagenesis was performed at 96-well scales using site-specific mutagenesis primers and full-length human ORF templates. PCR product was digested overnight using DpnI (NEB) without a ligation step to maximize throughput then transformed directly into competent cells to isolate single colonies. Four colonies per mutagenesis reaction were then hand-picked and arrayed into 96-well plates. After 21 hrs incubation at 37°C, glycerol stocks were generated then clones were pooled into four respective bacterial pools. Maxiprepped DNAs from each of the four pools were then combined through multiplexing (NEBNext) then sequenced in a single 1x100 single-end Illumina HiSeq run. Properly mutated clones were then identified by next-generation sequencing analysis and recovered from single-colony glycerol stocks. In total, we generated individual clones for 208 dnMis mutations comprising 109 from ASD probands and

99 from unaffected siblings.

4.6.3 Experimental examination of 667 protein-protein interactions using our high-throughput yeast two-hybrid (Y2H) assay

To perform Y2H, pDEST-AD and pDEST-DB plasmid vectors corresponding to the GAL4 activating domain (AD) and DNA-binding (DB) domain, respectively, were used. Full-length Clone-seq identified mutant clones were transferred into Y2H-amenable pDEST-DB and pDEST-AD vectors by Gateway LR reactions then transformed into *MAT*α Y8930 and *MAT*a Y8800, respectively. All DB-ORF *MAT*α transformants, including wild-type ORFs, were then mated against corresponding wild-type (WT) and mutant AD-ORF *MAT*a transformants in a pairwise orientation on YEPD agar plates. After mating, yeasts were replica-plated onto selective SC-Leu-Trp-His+ 1 mM of 3-amino-1,2,4-triazole (3AT) as well as SC-Leu-Trp-Adenine plates. Interactions were scored after 3 days of incubation and 5 days of incubation for SC-Leu-Trp+3AT and SC-Leu-Trp-Ade plates, respectively. To screen out autoactivating DB-ORFs, all DB-ORF *MAT*α transformants are also mated pairwise against empty pDEST-AD *MAT*a transformants and scored for growth on SC-Leu-Trp+3AT and SC-Leu-Trp-Ade plates. DB-ORFs that trigger reporter activity under this setup are removed from further experiments. We finally examined 667 interactions, of which the WT proteins could be detected with strong Y2H-positive phenotypes in our experiments, for 151 out of the 208 total dnMis mutations that we have successfully generated. The other 57 dnMis mutations corresponded to proteins with no testable interaction partners by Y2H; therefore, they were excluded from Y2H

experiments. While on average each of the 151 mutations was tested against 4-5 interaction partners, two proband mutations (Q8TBB1:p.Glu295Lys and Q8TD31:p.Trp337Arg) had >40 interaction partners tested and disrupted >30 of their corresponding interactions. Thus we excluded these two outliers when comparing the disruption rates of dnMis mutations in ASD probands and unaffected siblings (**Figure 4.4-2A**).

4.6.4 Computational prediction for protein-protein interaction disruption

For the remaining 1,582 dnMis mutations, we assessed their probabilities to disrupt an interaction based on whether they are likely to be on protein interaction interfaces and whether they tend to have damaging functional effects on the protein. We first applied an ensemble machine learning algorithm to predict interface residues (Interactome INSIDER). For each of these dnMis mutations, on each of its interactions with an interaction specific partner, we considered a mutation to be an interaction interface residue for this specific interaction if it has a probability score of very high, high or medium in Interactome INSIDER prediction. We next evaluated its deleteriousness to the protein using PolyPhen-2 (PPH2). If a mutation predicted as an interface residue also has a "probably damaging" PPH2 score (Interface+ and PPH2+), we considered this mutation to disrupt the interaction. On the other hand, we called a mutation non-disrupting if it was predicted to be unlikely an interaction interface residue (probability below "medium" by Interactome INSIDER) and to be "benign" to the protein by PPH2 (Interface- and PPH2-). Considering that using individual measurements (PPH2 alone or Interactome INSIDER alone) does not provide

sufficient signal towards whether a mutation is damaging or not, mutations that only meet one of these two criteria (Interface+ and PPH2-; Interface- and PPH2+) were excluded from the analyses. Importantly, when we included all the Interface+PPH2- and Interface-PPH2+ mutations as non-disrupting to our analyses, we found that all our results remain the same.

4.6.5 Modeling the number of disrupted interactions as a function of case-control status

Some missense mutations fail to disrupt any interactions, $D = 0$ disruptions. Other mutations, however, can disrupt $D = 1, 2, \ldots, I$ interactions. To account for the dispersion in D, and to determine if D was stochastically greater for missense mutations found in ASD probands versus unaffected siblings, we modeled D as a negative binomial distribution and fit it to case-control status. We also evaluated other models for goodness-of-fit, specifically Poisson and zero-inflated versions of Poisson and negative binomial. After accounting for degrees of freedom, none of these models fit the data as well as the negative binomial by the Akaike information criterion.

4.6.6 Construction of plasmids for Western blot and co-immunoprecipitation

Wild-type RARA and RXRB entry clones were obtained from the hORFeome v8.1 (Yang et al., 2011) collection. Gateway LR reactions were used to transfer bait RARA wild-type, p.Pro375Leu, and p.Arg83His into a pQXIP (ClonTech, 631516) vector modified to include a Gateway cassette featuring a C-terminal 3×FLAG. Prey RXRB was transferred into pcDNA-DEST40 which includes a V5 tag (Invitrogen, 12274-015) also using Gateway LR reactions.

4.6.7 Cell culture, co-immunoprecipitation, and Western blotting

HEK 293T cells were maintained in complete DMEM medium supplemented with 10% FBS. Cells were grown in 6-well dishes to 70-80% confluency then transfected using 1 μg bait construct and 1ug prey construct with 10 μL of 1mg/mL PEI (Polysciences Inc, 23966) mixed thoroughly with 150 μL OptiMEM (Gibco, 31985-062). After 24 hrs incubation, cells were gently washed three times in 1x PBS and then resuspended in 200 μL cell lysis buffer [10 mM Tris-Cl pH 8.0, 137mM NaCl, 1% Triton X-100, 10% glycerol, 2 mM EDTA, and 1x EDTA-free Complete Protease Inhibitor tablet (Roche)] and incubated on ice for 30 min. Extracts were cleared by centrifugation for 10 mins at 13,000 rpm at 4℃. For co- immunoprecipitation, 100 μL cell lysate per sample were incubated with 5 μL EZ view Red Anti-FLAG M2 Affinity Gel (Sigma, F2426) for 2 hrs at 4°C under gentle rotation. After incubation, bound proteins were washed three times in cell lysis buffer then eluted in 50 μL elution buffer (10 mM Tris-Cl pH 8.0, 1% SDS) at 65°C for 10 min. Cell lysates and co-immunoprecipitated samples were then treated in 6x SDS protein loading buffer (10% SDS, 1 M Tris-Cl pH 6.8, 50% glycerol, 10% β-mercaptoethanol, 0.03% Bromophenol blue) and subjected to SDS-PAGE. Proteins were then transferred from gels onto PVDF (Amersham) membranes. Anti-FLAG (Sigma, F1804), anti-V5 (Invitrogen, R960-25), and anti-γ-Tubulin (Sigma, T5192) at 1:5000, 1:3000, and 1:3000 dilutions, respectively, were used for immunoblotting analysis.

4.6.8 Evaluation of the distance between gene sets in the interactome network

We evaluated the distance between two gene sets using the method previously

published by Neale *et al.* (Neale et al., 2012b):in an interactome background, the distance between two gene sets (L_1 and L_2) is the average distance of each gene i in L_1 to L_2, where the distance of a specific gene i in L_1 to $L_2$is the average distance of gene i to each gene j in L_2. Let n_1 and n_2 be the number of genes in L_1 and L_2,

$$Distance(L_1, L_2) = \frac{1}{n_1}\sum_i Distance(i, L_2)$$

where $Distance(i, L_2) = \frac{1}{n_2}\sum_j Distance(i, j)$.

Then consider i and j as two nodes in the interactome network, the distance between these two nodes [$Distance(i, j)$] here is defined as the minimum number of intermediate nodes that connect i and j in the shortest path.

4.7 References

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

Albert, R., Jeong, H., and Barabasi, A.L. (2000). Error and attack tolerance of complex networks. Nature *406*, 378-382.

Basu, S.N., Kollu, R., and Banerjee-Basu, S. (2009). AutDB: a gene reference resource for autism research. Nucleic acids research *37*, D832-836.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res *28*, 235-242.

Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J.M., Murray, R.R., Roncari, L., de Smet, A.S.*, et al.* (2009). An experimentally derived confidence score for binary protein-protein interactions. Nat Methods *6*, 91-97.

Bruneau, B.G. (2008). The developmental genetics of congenital heart disease. Nature *451*, 943-948.

Chang, J., Gilman, S.R., Chiang, A.H., Sanders, S.J., and Vitkup, D. (2015). Genotype to phenotype relationships in autism spectrum disorders. Nat Neurosci *18*, 191-198.

Chatr-Aryamontri, A., Breitkreutz, B.J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L.*, et al.* (2015). The BioGRID interaction database: 2015 update. Nucleic Acids Res *43*, D470-478.

Chen, W.H., Lu, G., Chen, X., Zhao, X.M., and Bork, P. (2017). OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. Nucleic Acids Res *45*, D940-D944.

Das, J., and Yu, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Systems Biology *6*, 92.

de Ligt, J., Veltman, J.A., and Vissers, L.E. (2013). Point mutations as a source of de novo genetic disease. Curr Opin Genet Dev *23*, 257-263.

de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C.*, et al.* (2012). Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med *367*, 1921-1929.

De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S.*, et al.* (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature *515*, 209-215.

Deciphering Developmental Disorders, S. (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature *519*, 223-228.

Deciphering Developmental Disorders, S. (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature *542*, 433-438.

Devlin, B., and Scherer, S.W. (2012). Genetic architecture in autism spectrum disorder. Curr Opin Genet Dev *22*, 229-237.

Dong, S., Walker, M.F., Carriero, N.J., DiCola, M., Willsey, A.J., Ye, A.Y., Waqar, Z., Gonzalez, L.E., Overton, J.D., Frahm, S.*, et al.* (2014). De novo insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. Cell Rep *9*, 16-23.

Epi, K.C., Epilepsy Phenome/Genome, P., Allen, A.S., Berkovic, S.F., Cossette, P., Delanty, N., Dlugos, D., Eichler, E.E., Epstein, M.P., Glauser, T.*, et al.* (2013). De novo mutations in epileptic encephalopathies. Nature *501*, 217-221.

Euro, E.-R.E.S.C., Epilepsy Phenome/Genome, P., and Epi, K.C. (2014). De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. Am J Hum Genet *95*, 360-370.

Feldman, I., Rzhetsky, A., and Vitkup, D. (2008). Network properties of genes harboring inherited disease mutations. Proc Natl Acad Sci U S A *105*, 4323-4328.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A.*, et al.* (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res *44*, D279-285.

Fischbach, G.D., and Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron *68*, 192-195.

Fromer, M., Pocklington, A.J., Kavanagh, D.H., Williams, H.J., Dwyer, S., Gormley, P., Georgieva, L., Rees, E., Palta, P., Ruderfer, D.M.*, et al.* (2014). De novo mutations in schizophrenia implicate synaptic networks. Nature *506*, 179-184.

Geisheker, M.R., Heymann, G., Wang, T., Coe, B.P., Turner, T.N., Stessman, H.A.F., Hoekzema, K., Kvarnung, M., Shaw, M., Friend, K.*, et al.* (2017). Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. Nat Neurosci *20*, 1043-1051.

Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A.*, et al.* (2014). Genome sequencing identifies major causes of severe intellectual disability. Nature *511*, 344-347.

Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.L. (2007). The human disease network. Proc Natl Acad Sci U S A *104*, 8685-8690.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A.*, et al.* (2004). IntAct: an open source molecular interaction database. Nucleic Acids Res *32*, D452-455.

Huang, N., Lee, I., Marcotte, E.M., and Hurles, M.E. (2010). Characterising and predicting haploinsufficiency in the human genome. PLoS Genet *6*, e1001154.

Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E.*, et al.* (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216-221.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.H., Narzisi, G., Leotta, A.*, et al.* (2012). De novo gene disruptions in children on the autistic spectrum. Neuron *74*, 285-299.

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

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A.*, et al.* (2009). Human Protein Reference Database--2009 update. Nucleic Acids Res *37*, D767-772.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet *46*, 310-315.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B.*, et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285-291.

Levy, D., Ronemus, M., Yamrom, B., Lee, Y.H., Leotta, A., Kendall, J., Marks, S., Lakshmi, B., Pai, D., Ye, K.*, et al.* (2011). Rare de novo and transmitted copy-number variation in autistic spectrum disorders. Neuron *70*, 886-897.

Mefford, H.C., Batshaw, M.L., and Hoffman, E.P. (2012). Genomics, intellectual disability, and autism. N Engl J Med *366*, 733-743.

Mewes, H.W., Ruepp, A., Theis, F., Rattei, T., Walter, M., Frishman, D., Suhre, K., Spannagl, M., Mayer, K.F., Stumpflen, V.*, et al.* (2011). MIPS: curated databases and comprehensive secondary data resources in 2010. Nucleic Acids Res *39*, D220-224.

Meyer, M.J., Beltran, J.F., Liang, S., Fragoza, R., Rumack, A., Liang, J., Wei, X., and Yu, H. (2018). Interactome INSIDER: a structural interactome browser for genomic studies. Nat Methods.

Meyer, M.J., Das, J., Wang, X., and Yu, H. (2013). INstruct: a database of high-quality 3D structurally resolved protein interactome networks. Bioinformatics *29*, 1577-1579.

Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.F., Stevens, C., Wang, L.S., Makarov, V.*, et al.* (2012a). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature *485*, 242-245.

Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V.*, et al.* (2012b). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature *485*, 242.

O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D.*, et al.* (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246-250.

Pengelly, R.J., Greville-Heygate, S., Schmidt, S., Seaby, E.G., Jabalameli, M.R., Mehta, S.G., Parker, M.J., Goudie, D., Fagotto-Kaufmann, C., Mercer, C.*, et al.* (2016). Mutations specific to the Rac-GEF domain of TRIO cause intellectual disability and microcephaly. Journal of Medical Genetics *53*, 735-742.

Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S.*, et al.* (2010). Functional impact of global rare copy number variation in autism spectrum disorders. Nature *466*, 368-372.

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

Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S.E., Kahler, A.*, et al.* (2014). A polygenic burden of rare disruptive mutations in schizophrenia. Nature *506*, 185-190.

Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di Donato, N.*, et al.* (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet *380*, 1674-1682.

Reijnders, M.R.F., Ansor, N.M., Kousi, M., Yue, W.W., Tan, P.L., Clarkson, K., Clayton-Smith, J., Corning, K., Jones, J.R., Lam, W.W.K.*, et al.* (2017). RAC1 Missense Mutations in Developmental Disorders with Diverse Phenotypes. The American Journal of Human Genetics *101*, 466-477.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, Samuel J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R.*, et al.* (2014). A Proteome-Scale Map of the Human Interactome Network. Cell *159*, 1212-1226.

Ronemus, M., Iossifov, I., Levy, D., and Wigler, M. (2014). The role of de novo mutations in the genetics of autism spectrum disorders. Nat Rev Genet *15*, 133-141.

Ropers, H.H. (2010). Genetics of early onset cognitive impairment. Annu Rev Genomics Hum Genet *11*, 161-187.

Sadybekov, A., Tian, C., Arnesano, C., Katritch, V., and Herring, B.E. (2017). An autism spectrum disorder-related de novo mutation hotspot discovered in the GEF1 domain of Trio. Nature Communications *8*, 601.

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J.I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G.I., Wang, Y.*, et al.* (2015). Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders. Cell *161*, 647-660.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. Nucleic Acids Res *32*, D449-451.

Sanders, S.J., Ercan-Sencicek, A.G., Hus, V., Luo, R., Murtha, M.T., Moreno-De-Luca, D., Chu, S.H., Moreau, M.P., Gupta, A.R., Thomson, S.A.*, et al.* (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron *70*, 863-885.

Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S.*, et al.* (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron *87*, 1215-1233.

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L.*, et al.* (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237-241.

Schanen, N.C. (2006). Epigenetics of autism spectrum disorders. Hum Mol Genet *15 Spec No 2*, R138-150.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J.*, et al.* (2007). Strong association of de novo copy number mutations with autism. Science *316*, 445-449.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S.*, et al.* (2005). A human protein-protein interaction network: a resource for annotating the proteome. Cell *122*, 957-968.

Turner, B., Razick, S., Turinsky, A.L., Vlasblom, J., Crowdy, E.K., Cho, E., Morrison, K., Donaldson, I.M., and Wodak, S.J. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. Database (Oxford) *2010*, baq023.

Turner, T.N., Yi, Q., Krumm, N., Huddleston, J., Hoekzema, K., HA, F.S., Doebley, A.L., Bernier, R.A., Nickerson, D.A., and Eichler, E.E. (2017). denovo-db: a compendium of human de novo variants. Nucleic Acids Res *45*, D804-D811.

Venkatesan, K., Rual, J.F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.I.*, et al.* (2009). An empirical framework for binary interactome mapping. Nat Methods *6*, 83-90.

Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander, E.S., and Sabatini, D.M. (2015). Identification and characterization of essential genes in the human genome. Science *350*, 1096-1101.

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol *30*, 159-164.

Wei, X., Das, J., Fragoza, R., Liang, J., Bastos de Oliveira, F.M., Lee, H.R., Wang, X., Mort, M., Stenson, P.D., Cooper, D.N.*, et al.* (2014). A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. PLoS Genet *10*, e1004819.

Xu, J., and Li, Y. (2006). Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics *22*, 2800-2805.

Xu, X., Li, C., Gao, X., Xia, K., Guo, H., Li, Y., Hao, Z., Zhang, L., Gao, D., Xu, C.*, et al.* (2017). Excessive UBE3A dosage impairs retinoic acid signaling and synaptic plasticity in autism spectrum disorders. Cell Research.

Yang, X., Boehm, J.S., Yang, X., Salehi-Ashtiani, K., Hao, T., Shen, Y., Lubonja, R., Thomas, S.R., Alkan, O., Bhimdi, T.*, et al.* (2011). A public genome-scale lentiviral expression library of human ORFs. Nat Methods *8*, 659-661.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N.*, et al.* (2008). High-quality binary protein interaction map of the yeast interactome network. Science *322*, 104-110.

Zaidi, S., Choi, M., Wakimoto, H., Ma, L., Jiang, J., Overton, J.D., Romano-Adesman, A., Bjornson, R.D., Breitbart, R.E., Brown, K.K.*, et al.* (2013). De novo mutations in histone-modifying genes in congenital heart disease. Nature *498*, 220-223.

CHAPTER 5

SUMMARY AND FUTURE DIRECTIONS

5.1 Per-chapter summary

This dissertation has presented several interactome-based approaches towards interrogating the function of missense disease-associated mutations (Chapter 2), population variants (Chapter 3), and *de novo* mutations (Chapter 4). To frame these efforts within the context of human population genetics, Chapter 1 began by familiarizing the reader with fundamental concepts regarding random genetic drift and selection. While genetic drift and selection are sophisticated and extensive topics, straightforward derivations and simulations of Wright-Fisher populations reveal that random genetic drift can dominate the allele frequency behavior of weakly deleterious, *de novo* mutations, particularly within shorter time frames. Consequently, weakly deleterious alleles are expected to be widespread in human populations, particularly those under rapid population growth.

Following this brief primer on population genetics, a short review of human interactome networks and disease was presented. Using a protein interaction framework, genotype-to-phenotype connections can instead be interpreted as specific perturbations to edges in protein-protein interaction networks. Structurally-resolved interactome networks have increased the resolution of these approaches and an example of gene pleiotropy in the gene *WASP*, in which mutations to distinct interaction interfaces of WASP result in clinically distinct diseases, was presented. Considering the limited success of functional prediction algorithms, such as PolyPhen-

2, in generating discernible phenotypes in mice, systematic profiling of protein interaction perturbations by Y2H may represent an alternative and more informative approach towards characterizing functional coding variation in human genomes.

Chapter 1 ended with an extensive discussion focused principally on seven high profile, whole-genome and whole-exome sequencing efforts across different populations (summarized in **Table 1.6-1**). In brief, the 1000 Genomes Project, ESP6500, and ExAC sequencing efforts have all unveiled pervasive rare variation segregating across human populations as a result of recent explosive population growth. The Phase I iterations of the 1000 Genomes Project and ESP6500 go on to estimate that up to 2.3% of rare coding variants in human genomes are expected to impact protein function. The UK10K and Icelandic genome sequencing efforts go further by actually performing GWAS for the rare coding variation revealed in their studies. Outside of a handful of strong associations, though, little evidence in support of low-frequency alleles with large effects on traits was found. Lastly, efforts to characterize *de novo* variants across a single population and in an ASD cohort were detailed by the GoNL and SSC sequencing efforts, respectively. Both efforts demonstrated that the rate of *de novo* mutation occurrence depends directly on the parents' age while the ASD effort showed that 43% of proband likely-gene-damaging mutations contribute to ASD diagnoses in comparison to only 13% for missense probands.

In Chapter 2, Clone-seq, a novel method for cloning DNA coding variants on a massively parallel scale, was described. Clone-seq was then used to generate and study the impact of 204 disease-associated mutations on protein interactions and

stability. A Y2H approach revealed that disease-associated mutations on protein-protein interaction interfaces disrupted corresponding interactions in 78% of cases while this disruption rate dropped to 29% when mutations occurred away from interaction interfaces. Moreover, mutation pairs on the same gene that disrupted the same set of interactions were more likely to map back to the same disease in comparison to when different interaction partners were disrupted. Similarly, mutation pairs in which both mutant proteins were either stably expressed or both were unstable were also more likely to map back to the same disease in comparison to cases in which one mutation was stable and the other was not.

Lastly, a case study of two SMAD4 mutations on the interaction interface with SMAD3 as well as a third SMAD4 mutation away from the interaction interface was presented. Accordingly, Y2H revealed that the two equally disruptive mutations on the interaction interface resulted in the same disease, juvenile polyposis coli, while the non-disruptive SMAD4 mutation away from the interaction interface resulted in a clinically distinct disease, pulmonary arterial hypertension. This SMAD4 example was particularly compelling because it begged the question of whether a rare population variant that disrupts the same set of interactions as a known disease-associated mutation could result in the same disease phenotype – a topic further explored in Chapter 3.

Chapter 3 applied Clone-seq to a study of >2,000 missense SNVs consisting predominantly of population variants from ExAC but also of disease-associated mutations from HGMD and cancer somatic mutations from COSMIC. A total of 309 disruptive variants from ExAC were identified and their disruption rate was found to

vary inversely with respect to allele frequency. When the site frequency spectrum from ExAC was combined with these measured disruption rates, 11.2% of missense variants per individual genome were calculated to be disruptive, significantly higher than the 2.3% figure presented by the ESP6500 Phase I study. Moreover, disruptive variants were seldom the consequence of unstable protein folding, implying that most missense variants result in local structural perturbations that disrupt specific protein interactions as opposed to destabilizing protein stability as a whole. Disruptive variants were also found to be enriched on conserved genomic sites, more likely to be population-specific, and enriched within genomic regions under positive selection.

Despite an unexpectedly high fraction of disruptive alleles found, the impact of these disruptive mutations appears to be readily mitigated at gene and protein interaction network levels. As a result, disruptive variants were found to be heavily depleted in essential genes, including oncogenes and haploinsufficient genes. Moreover, null-like variants, which disrupt all tested interaction partners, had significantly lower betweenness centrality values, further indicating that disruptive variants principally target non-essential genes. Lastly, by comparing the interaction perturbation profiles of disruptive population variants with those of disease-associated mutations, a candidate disease-associated variant, T152I, on the enzyme PSPH was identified. A malachite green assay then revealed that PSPH T152I reduced enzymatic activity to an equal extent as that of PSPH D32N, suggesting that the rare variant T152I may phenocopy the disease mutation D32N in a compound heterozygous background.

Finally, in Chapter 4 interactome-scale protein structural data was combined

with deleteriousness scores from PolyPhen-2 to construct a computational pipeline for prioritizing *de novo* missense mutations that contribute to disease phenotypes. The study took advantage of whole-genome sequencing data for >2,500 families with an ASD-afflicted child and unaffected parents and siblings. dnMis mutations in probands were found to be enriched on interaction interfaces in comparison to dnMis mutations in unafflicted siblings, suggesting that proband mutations frequently result in protein interaction perturbations. Accordingly, a Y2H study of 151 cloned dnMis mutations across 667 mutation-interaction pairs unveiled that ASD proband mutations were significantly more likely to be disruptive.

To achieve interactome scales, a two-tiered prediction approach was then developed in which dnMis mutations that (1) occurred on high confidence interaction interface residues and (2) were predicted to be "probably damaging" by PolyPhen-2 were scored as *interaction-disruptive* while dnMis mutations that failed to meet either criteria were scored as *interaction non-disruptive*. Impressively, this two-tiered prediction approach achieved ~80% accuracy when tested against our subset of 667 Y2H-scored mutation-interaction pairs. Moreover, this computational approach identified a proband dnMis mutation on the gene-encoded protein, RARA, predicted to occur on an interaction interface residue with RXRB. An RARA dnMis mutation on an unaffected sibling, in contrast, occurred away from any interaction interface residues and was therefore predicted to be non-disruptive. Both Y2H and co-IP confirmed that the proband dnMis mutation disrupted the RARA-RXRB interaction while the unaffected sibling dnMis mutation left this interaction intact, in agreement with our predictions.

5.2 Future directions

The turn of the century was marked by the completion of the human genome project. The project was celebrated as a pivotal step towards a personal genomics revolution. Nearly a decade later, a thousand genomes were fully sequenced. Only three to four years after that, a thousand genomes turned into ~10,000 exomes as part of the UK10K project followed shortly by >60,000 exomes as part of ExAC. A beta version of a database named gnomAD is now available consisting of >120,000 fully sequenced exomes as well as whole-genome sequencing data for an additional 20,000 individuals. The sequencing revolution is in full effect and the era of personal genomics has already entered advanced phases. Sequencing genomes is no longer a road block. Rather, the rate-limiting step is properly interpreting this deluge of sequencing data. The interactome-based studies presented in this thesis are an important step towards proper interpretation of genomic variation. By understanding the biophysical and evolutionary properties of interaction-disruptive variants in comparison to non-disruptive variants, better predictors for functional, disease-relevant variation can be achieved. This is a crucial development since ultimately a disease-relevant functional prediction algorithm is needed to keep pace with the scale of variant discovery. Continuing high-throughput experimental studies focused not just on coding variation, but on widespread non-coding variation as well, will serve as indispensible resources towards our constant pursuit of a compete understanding of functional, medically-actionable variants segregating across human populations.

APPENDIX A

SUPPORTING INFORMATION FOR CHAPTER 1

A.1 *Assumptions in Wright Fisher Model and Effective Population Sizes*

Real-world populations are far from ideal and therefore complicate population genetic calculations. Consequently, an ideally behaving population is assumed in the Wright-Fisher model in which each generation is discrete. An individual in one generation is therefore not found in any other generation and a proceeding generation completely replaces its preceding generation. Constant population size, N, is not required but for convenience, constant population size is always assumed in calculations in this thesis. In a Wright-Fisher Model subject only to random genetic drift, parents from a preceding generation are chosen by random sampling with replacement and can therefore be modeled by a binomial distribution. Order of individuals does not matter, and the allele count, assuming random genetic drift, is constant per generation.

Any population that is assumed to behave ideally is referred to as an effective population with effective population size, N_e. For the purposes of this thesis, any references to an effective population assume that the population follows the Wright-Fisher assumptions described here, though numerous other definitions for effective population size exist.

A.2 *Deriving Equation 1.2-8 for logistic growth*

To derive Equation 1.2-5 in the text, we begin by substituting Equation 1.2-1 into 1.2-

4 as shown and solve for the expected frequency of allele A at $t+1$:

$$E[x(t+1)] = \frac{\omega_A N_A(t)}{\omega_A N_A(t) + \omega_a N_a(t)} \tag{A.2-1}$$

This resulting equation is then rewritten as follows, noting that the allele count N_A divided by $N_A + N_a$ is equal to allele frequency x:

$$\begin{aligned} &= \frac{\omega_A N_A(t)}{\omega_A N_A(t) + \omega_a N_a(t)} \cdot \frac{1/\omega_a(N_A + N_a)}{1/\omega_a(N_A + N_a)} \\ &= \frac{\omega_A/\omega_a \cdot x}{\omega_A/\omega_a \cdot x + (1-x)} \end{aligned} \tag{A.2-2}$$

We now substitute Equation 1.2-2 into A.2-2 which yields the following:

$$= \frac{(1+s)\cdot x}{(1+s)\cdot x + (1-x)} \tag{A.2-3}$$

where the label s is used for convenience. The expected change of frequency between a generation is then calculated as follows:

$$\begin{aligned} E[\Delta x] &= E[x(t+1)] - x(t) \\ &= \frac{(1+s)\cdot x}{(1+s)\cdot x + (1-x)} - x \\ &= \frac{sx\cdot(1-x)}{1+sx} \\ E[\Delta x] &\approx sx\cdot(1-x), when\ s \ll 1 \end{aligned} \tag{A.2-4}$$

where the last step in the derivation holds the value of the selection coefficient s is very small. We recognize that the differential equation shown in the last step of A.2-4 is known as a logistic equation with a known solution provided below:

$$x(t) = \frac{1}{1 + \left(\frac{1}{x_0} - 1\right) e^{-st}} \tag{A.2-5}$$

Equation A.2-5 rewritten then matches Equation 1.2-8, shown below:

$$x(t) = \frac{x_0}{x_0 + (1 - x_0)e^{-st}} \tag{5.2-8}$$

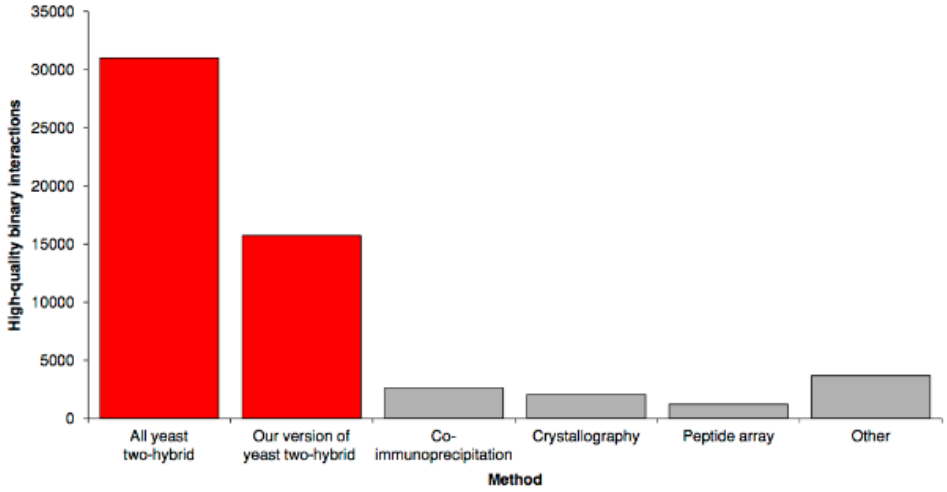
A.3 *Comment on the use of SNPs vs SNVs*

Briefly, the difference between a single nucleotide polymorphism, SNP, and single nucleotide variant, SNV, is often semantic and depends largely on the literature in which the term is used. While a SNV has no restriction on the type of point mutation that it can describe, including population variants, somatic mutations, and disease-associated mutations, the term SNP is typically reserved for studies exclusively regarding single nucleotide population variants. Consequently, the term SNV is used in Chapter 3 since three classes of variants (disease-associated, population, and somatic) are discussed. For simplicity consider the terms SNPs and SNVs interchangeable throughout Chapter 1.

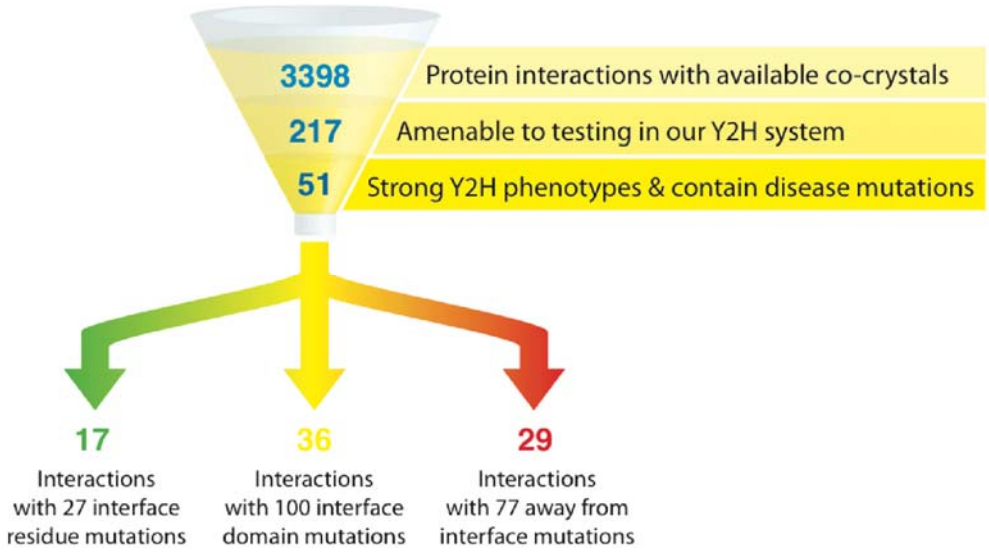
APPENDIX B

SUPPORTING INFORMATION FOR CHAPTER 2

B.1 *Supplementary Figures for Chapter 2*



Supplementary Figure B.1-1 Number of high-quality binary protein-interactions is highest for yeast two-hybrid (Y2H) with 16,000 interactions available using our version of Y2H.



Supplementary Figure B.1-2 Schematic of steps used to select mutations and interactions for our comparative interactome scanning pipeline

B.2 *Supplementary Text*

B.2.1 Probability of obtaining the desired clone

If the probability of successfully obtaining the desired clone with a single colony is p, assuming independence between the colonies, the chance of not obtaining even one desired clone after picking n colonies is $(1-p)^n$. Thus, the probability of obtaining at least one correct clone for one mutagenesis attempt is:

$$P(n) = 1 - (1-p)^n$$

Based on our HiSeq results, we can estimate p using the average fraction of desired clones obtained. Since we sequenced four colonies for each of the 39 desired clones, we have a total of 156 samples. Out of the 156 samples, 125 contain the desired mutations. Thus, $p = 125/156 = 0.8$. Substituting appropriate values, we calculate the probability of obtaining at least one correct clone for one mutagenesis attempt after picking 4 colonies, $P(4) = 0.998$.

In our pipeline, even colonies for generating different mutations of the same gene can be put into the same pool, which can be easily distinguished computationally when processing the sequencing results. Confusion only arises upon pooling colonies for generating the same mutation with identical surrounding sequences in the same gene or between different genes. In this situation, we can only identify the correct clones if all of these mutations in the same pool are correct. However, out of 50,491 missense disease mutations in HGMD and 395,780 coding SNPs in dbSNP, only 340 (~0.08%) will cause such confusion in Clone-seq. The probability of obtaining the desired clones for k instances of the same mutation with identical surrounding

sequences in the same gene or between different genes is given by:

$$P(n,k) = 1 - (1 - p^k)^n$$

Thus, even if we were to have 3 undistinguishable mutations with identical surrounding sequences (i.e., $k = 3$), after picking 4 colonies for each mutagenesis attempt, we would still have a 94% chance to have at least one pool out of the four where all three mutations are correct, rendering the whole Clone-seq pipeline successful.

B.2.2 Scalability of Clone-seq

The primary determinant of the scalability of our Clone-seq pipeline is the read coverage for alleles that we generate using our high-throughput mutagenesis PCR protocol. The average coverage of reads for each of the 39 alleles in our Clone-seq results is > 2,500×. For our Cloneseq results, we only used ~40 million reads out of a total of ~125 million reads in a single lane of a 1×100 bp HiSeq run. So, if we use all 125 million reads for the 4 colonies, we can sequence 39×(125/40) alleles with > 2,500× coverage. However, to determine *S* to a least count of 1%, we only need 100× coverage. Since the separation between a successful mutagenesis attempt with the lowest *S* and an unsuccessful mutagenesis attempt with the highest *S* is 0.28, 100× coverage makes this separation > 25 times our least count. We further increase this separation to > 60 times our least count by requiring $S > 0.8$ for a mutagenesis attempt to be considered successful. 100× coverage is also sufficient for a conservative variant calling pipeline to identify additional mutations with high confidence. Thus, we can generate 39×(125/40)×(2,500/100) = 3,047 mutant alleles with a single lane of a

1×100 bp HiSeq run using the Clone-seq pipeline.

B.2.3 Costs of Sanger sequencing vs. Clone-seq

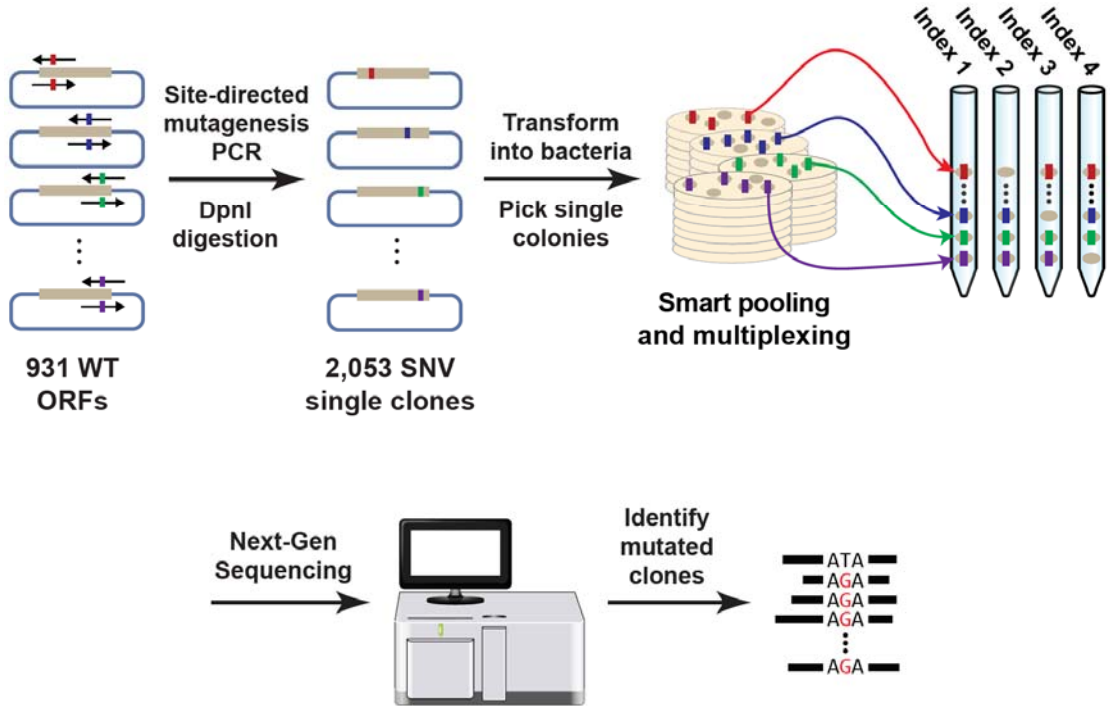
Table B.2-1 Cost comparison for Sanger vs Clone-seq. *Internal Cornell pricing applied.*

<table>
<tr><th colspan="2">Traditional Sanger sequencing</th><th colspan="2">Clone-seq</th></tr>
<tr><td>Unique mutations</td><td>3,047</td><td rowspan="2">NEBNext Multiplex Oligos (E7335S)</td><td rowspan="2">$19.80</td></tr>
<tr><td>Colonies per mutation</td><td>4</td></tr>
<tr><td>Total number of samples</td><td>3,047×4 = 12,188</td><td rowspan="2">NEBNext DNA Library Prep Master (E6040S)</td><td rowspan="2">$105</td></tr>
<tr><td>Re-sequencing needed[1]</td><td>5%</td></tr>
<tr><td>Number of 96-well plates needed</td><td>137</td><td rowspan="2">Illumina HiSeq, single-end, 100 bp sequencing lane</td><td rowspan="2">$1,175</td></tr>
<tr><td>Cost per plate</td><td>$300</td></tr>
<tr><td>Minimum cost[2]</td><td>43×$300 = $12,900</td><td rowspan="2">Total cost</td><td rowspan="2">$1,299.80</td></tr>
<tr><td>Total cost</td><td>137×$300 = $41,100</td></tr>
</table>

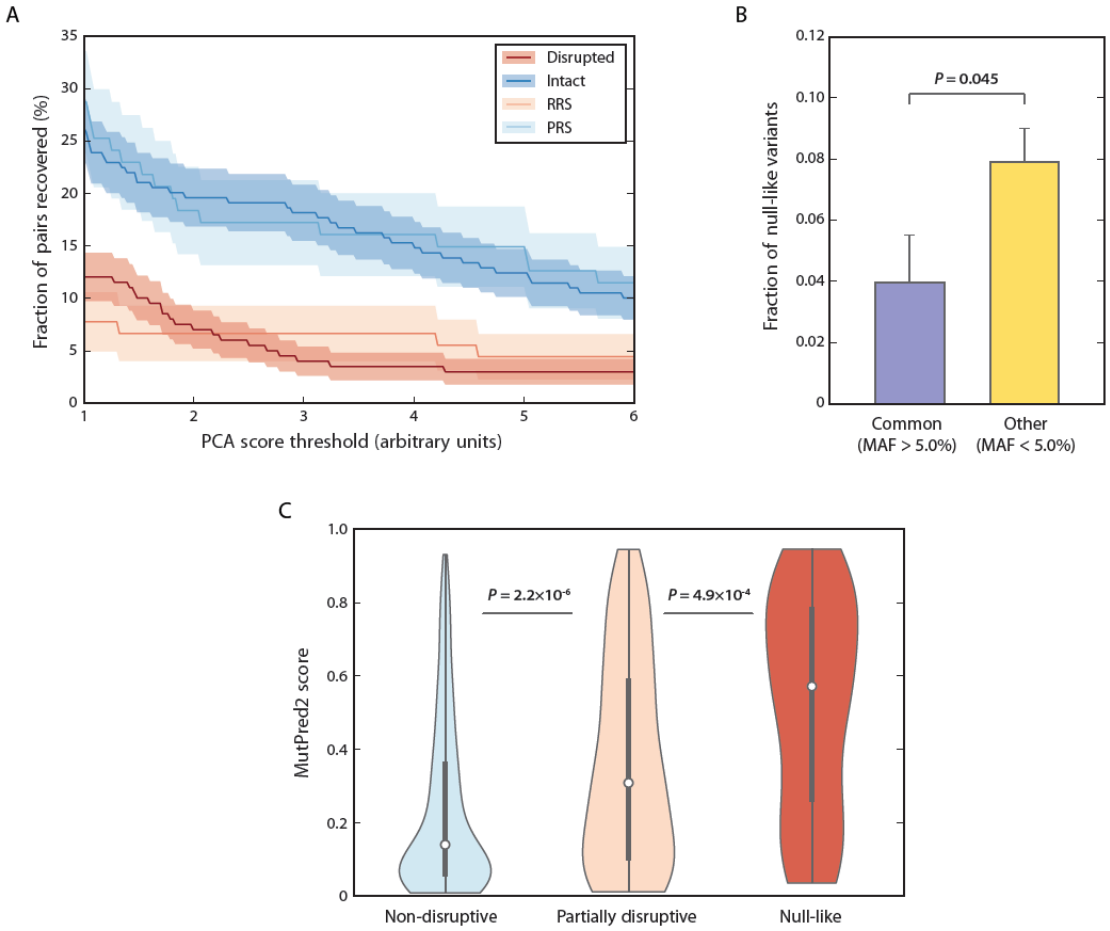
APPENDIX C

SUPPORTING INFORMATION FOR CHAPTER 3

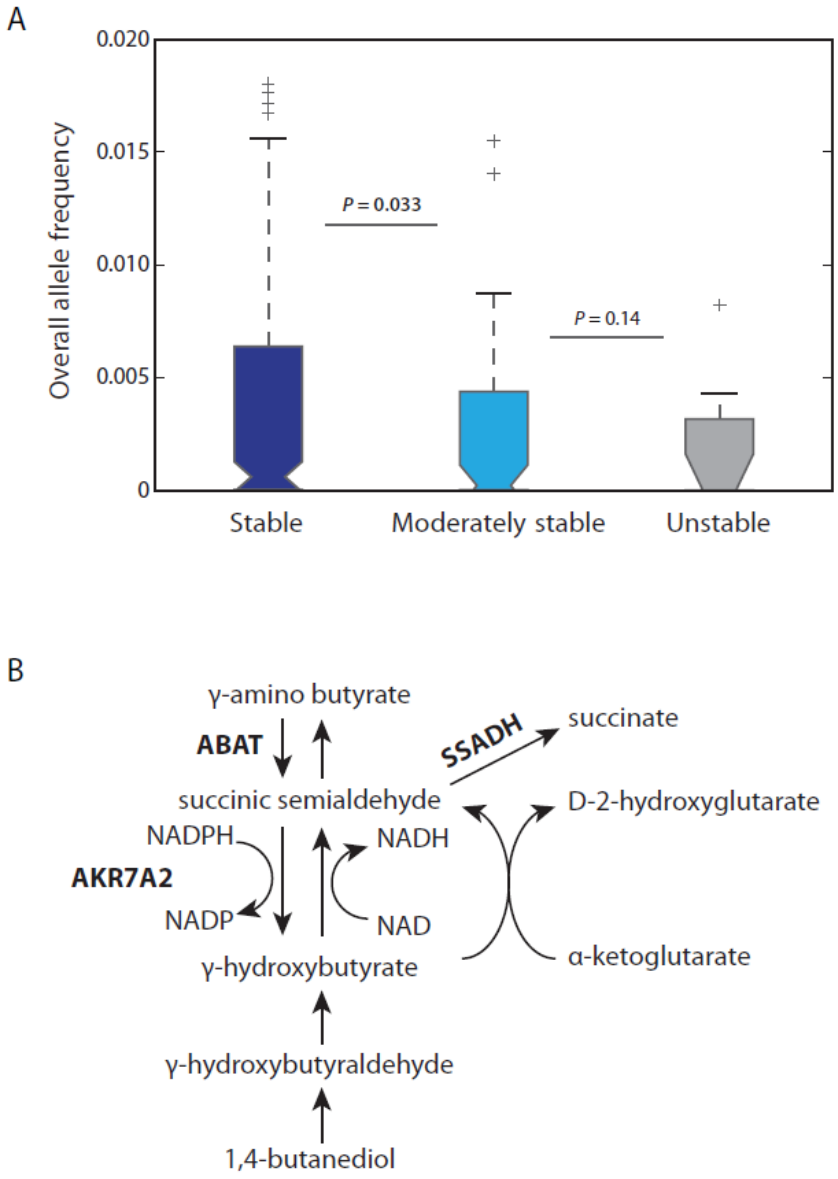
C.1 *Supplementary Figures for Chapter 3*



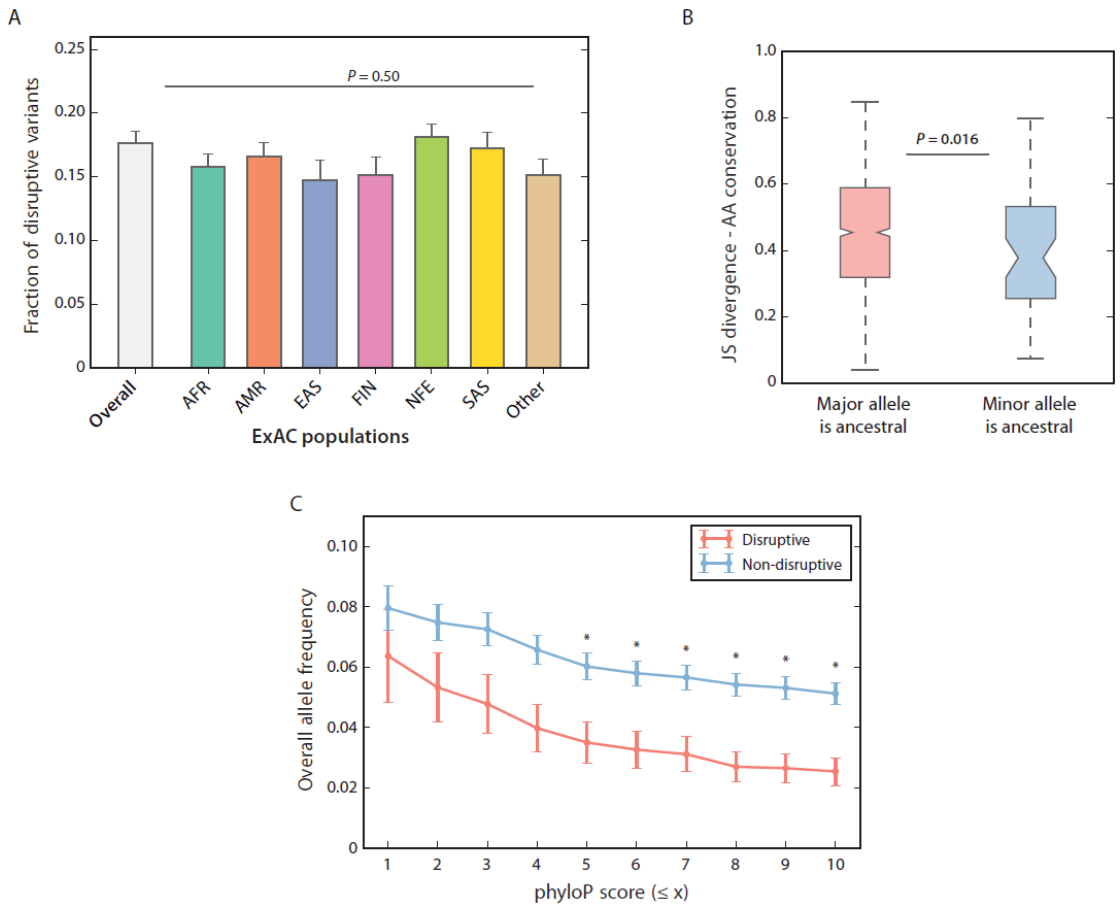
Supplementary Figure C.1-1 Clone-seq pipeline for generating 2,053 SNVs from 931 wild-type ORFs.



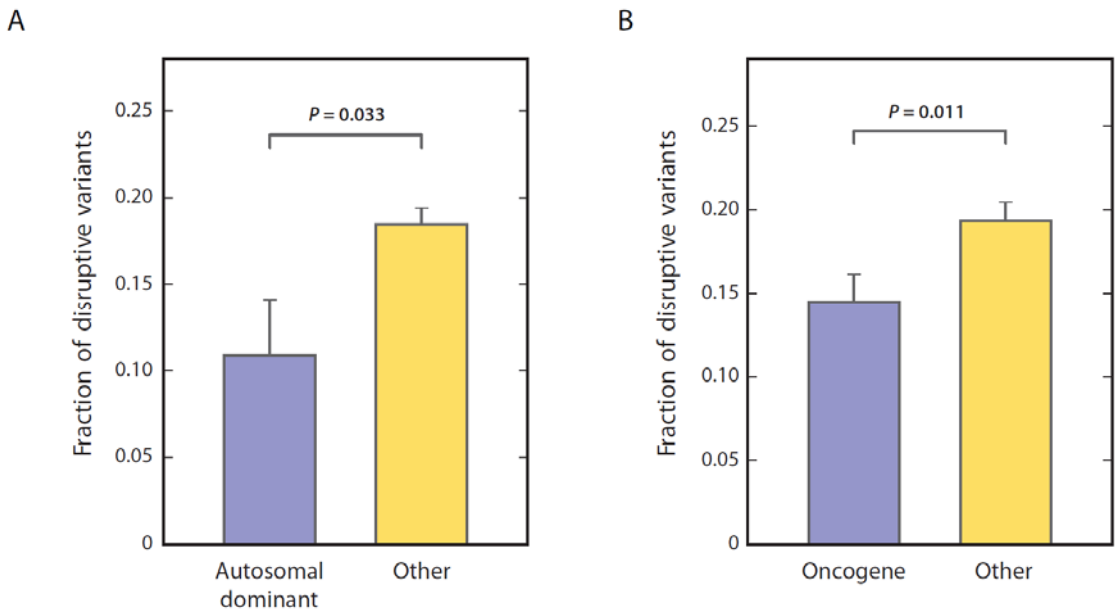
Supplementary Figure C.1-2 Null-like variants are depleted at common allele frequencies and are often predicted to be deleterious. *(A) Fraction of protein pairs recovered by PCA across increasingly stringent PCA scoring thresholds. SE of proportion is demarcated by shading. (B) Fraction of null-like variants that occur at minor allele frequency > 5.0% in comparison to those with minor allele frequency < 5.0%. Error bars indicate +SE of proportion. P value by one-tailed Z-test. (C) Distribution of MutPred2 scores across three disruption categories. Thick black bars are the interquartile range, white dots display the median, and extended thin black lines represent 95% confidence intervals. P values by one-tailed U-test. Significant P values in bold.*



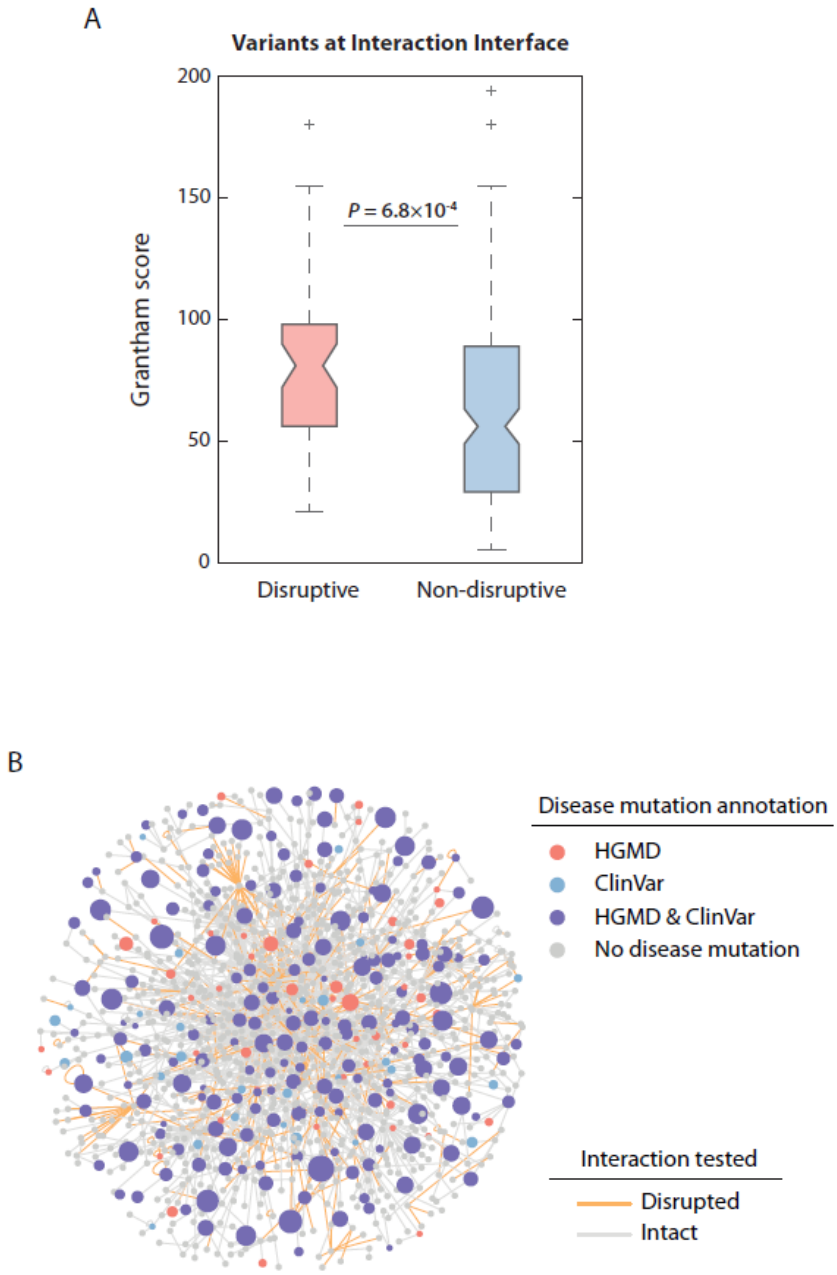
Supplementary Figure C.1-3 Variants that destabilize protein expression occur at rare allele frequencies. *(A) Distribution of overall allele frequencies for variants categorized as stable (n = 224), moderately stable (n = 57), and unstable (n = 10). P values by one-tailed U-test. Significant P values in* ***bold****. (B) γ-hydroxybutyrate metabolism pathways involving AKR7A2, ABAT, and SSADH in* ***bold****. Image is adapted from (Lyon et al., 2007).*



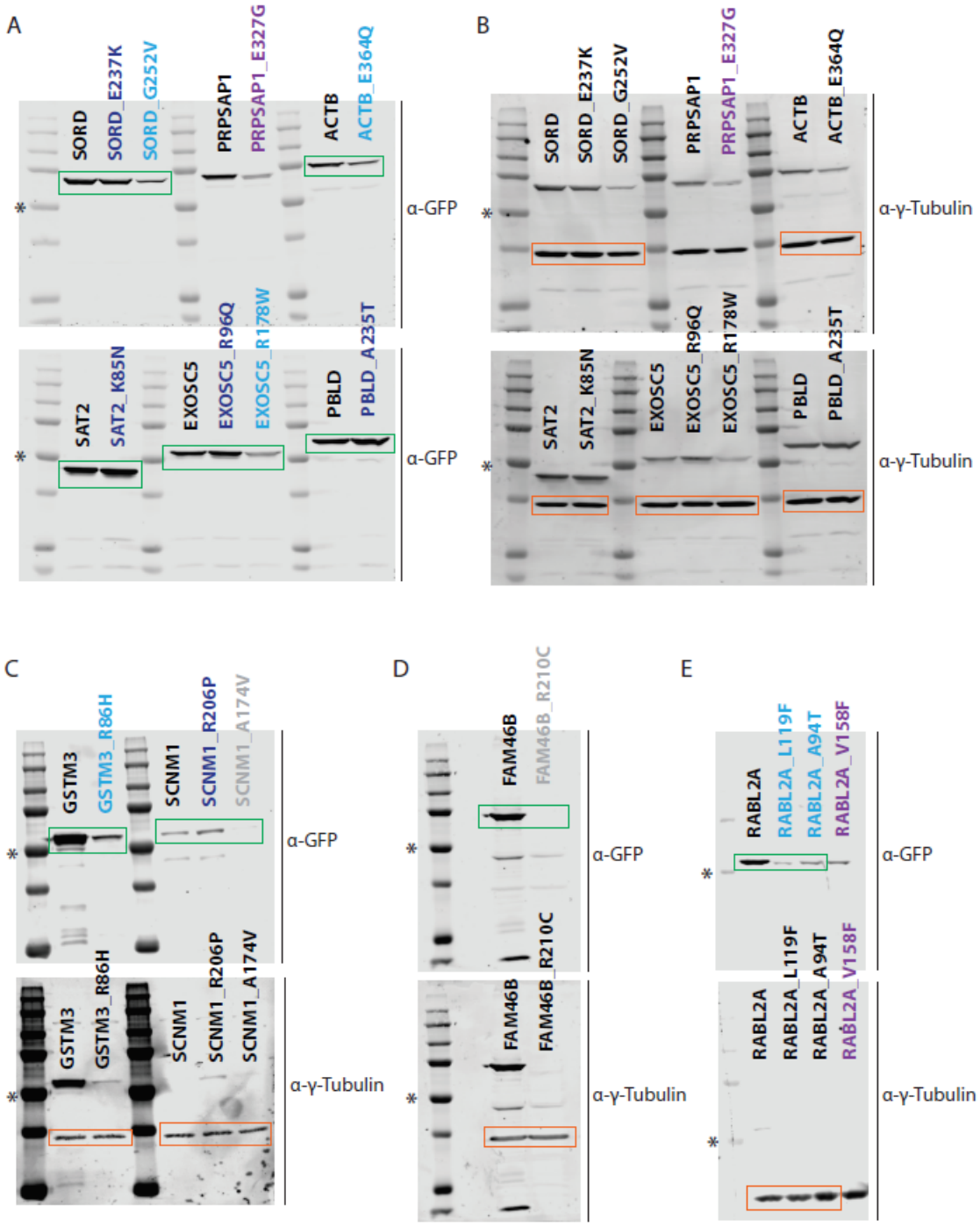
Supplementary Figure C.1-4 Disruptive alleles are not biased across specific ExAC populations. *(A) Fraction of disruptive variants across overall and individual ExAC populations. Error bars indicate +SE of proportion. P value by chi-square test. (B) Distribution of Jensen-Shannon Divergence scores for amino acid residues at sites in which the major allele equals the ancestral allele in comparison to sites in which the minor allele equals the ancestral allele. Larger scores indicate a more conserved site. P value by one-tailed U-test. (C) Relationship between conservation and overall allele frequency for disruptive and non-disruptive variants is examined across increasing phyloP cutoff scores. Error bars indicate ±SE of mean. P values by one-tailed Z-test. * $P < 0.05$.*



Supplementary Figure C.1-5 Disruptive variants are depleted among autosomal dominant disease-associated genes and known oncogenes. *(A) Fraction of disruptive variants on disease-associated genes with autosomal dominant inheritance pattern (n = 92) or other genes (n = 1,620). (B) Fraction of disruptive variants on oncogenes (n = 455) or other genes (n = 1,257). Error bars indicate +SE of proportion. P values by one-tailed Z-test.*

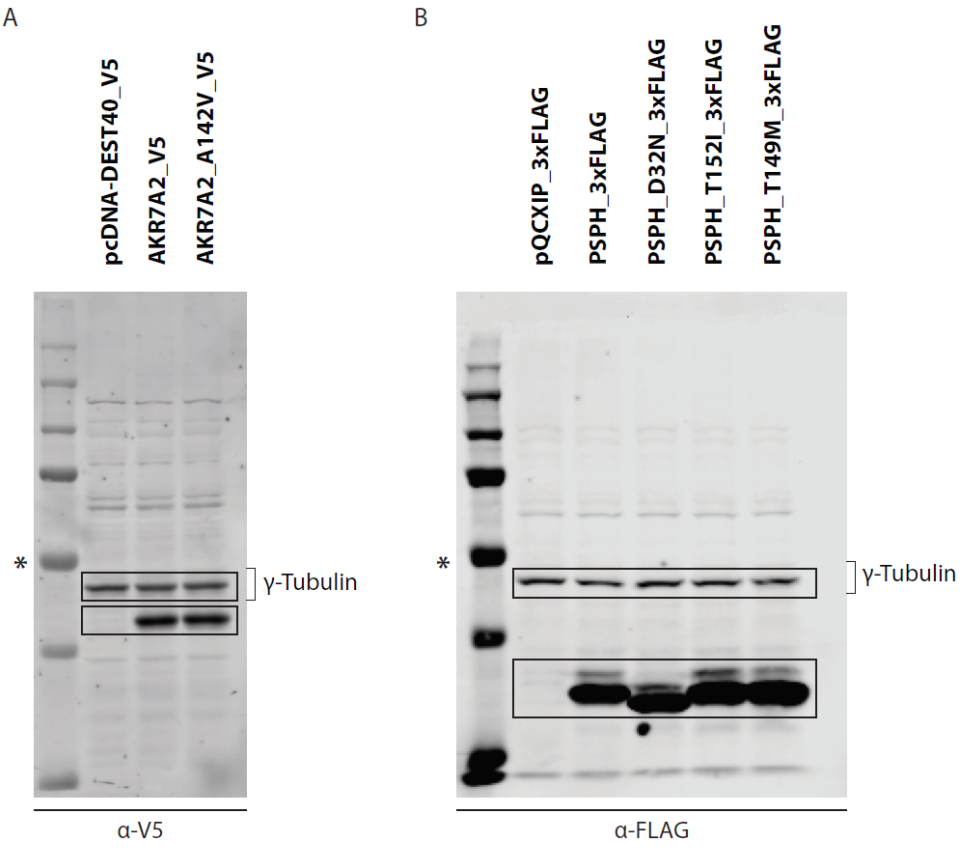


Supplementary Figure C.1-6 Candidate disease-associated mutations can be identified through shared interaction perturbation profiles. *(A) Distribution of Grantham scores for disruptive and non-disruptive variants on interaction interfaces for interactions with available co-crystal structures in PDB. P value by one-tailed U-test. (B) Protein-protein interaction network disruption by population variants in the context of human disease. Nodes represent a protein and edges represent Y2H-tested interactions. Interactions disrupted by a population variant are shown in orange. Node size is proportional to number of disease-associated mutations documented in HGMD and ClinVar per corresponding protein.*



Supplementary Figure C.1-7 Uncropped Western blots for stable, moderately stable, and unstable GFP expression examples in Figure 3B. *(A) Westerns for wild-type and corresponding mutant proteins detected by α-GFP. (B) α-GAPDH controls for westerns for wild-type and corresponding mutant proteins detected in A using stripped membrane. (C) Upper: Westerns for wild-type and corresponding mutant proteins detected by α-GFP. Lower:*

*α-GAPDH controls for western in upper using a stripped membrane. (D) Upper: Westerns for wild-type and corresponding mutant protein detected by α-GFP. Lower: α-GAPDH controls for western in upper. Membrane was not stripped. (E) Upper: Westerns for wild-type and corresponding mutant proteins detected by α-GFP. Lower: α-GAPDH controls for western in upper using a stripped membrane. In (A)-(E), stable, partially stable, and unstable mutations are labeled in blue, cyan, and gray, respectively. Partially stable mutations for population variants not listed in ExAC are shown in purple. Such mutations were not used in any analysis. Bands corresponding to α-GFP and α-GAPDH examples used in Figure 3B are encased in green and orange boxes, respectively. * indicates 50 kDa marker.*



Supplementary Figure C.1-8 Uncropped Western blots for AKR7A2 and PSPH mutant proteins. *(A) Westerns for wild-type and A142T variant of AKR7A2 detected by α-V5. α-γ-Tubulin control ran on an unstripped membrane. (B) Westerns for wild-type and mutant PSPH proteins detected by α-FLAG. α-γ-Tubulin control ran on an unstripped membrane. In (A) and (B), black boxes indicate where figures where cropped for Western blots in Figures 3G and 6G, respectively. * indicates 50 kDa marker.*

C.2 *Supplementary Text for Chapter 3*

C.2.1 Calculating the fraction of disruptive missense variants on a per-individual basis

We note that allele counts in ExAC correspond predominantly to alleles with MAF > 10%. Therefore the fraction of disruptive alleles that are common will have the greatest influence on the average number of interaction-disruptive variants per individuals. As such, we recalculated the disruption rates shown in **Figure 3.4-2B** for four MAF bins, listed in the table below, rather than three bins to increase the sensitivity of our calculation to common variants while still retaining robust across total allele counts.

MAF bins	<0.1%	0.1-1.0%	1-10%	>10%	Sum
Disruptive alleles	179	71	36	23	309
Total alleles	883	390	217	222	1712
Disruption rate	20.3%	18.2%	16.6%	10.4%	18.0%

Supplementary Table C.2-1 Disruption rate for tested ExAC alleles recalculated across four MAF bins.

Next, we calculated the site frequency spectrum for ExAC alleles annotated as *missense_variant* in at least one transcript by summing the adjusted overall allele counts (listed as *AC_adj* in the ExAC database) per MAF bin and dividing each bin count by the total adjusted overall allele count across all bins as shown in the formula below:

$$f_i = \frac{Allele\ count_i}{\sum_i^4 Allele\ count_i}$$

where i represents the four MAF bins examined and f_i represents the fraction of

missense variants per individual expected to be in MAF bin i. Applying this equation to all four MAF bins yields the following per-individual proportions:

MAF bins	<0.1%	0.1-1.0%	1-10%	>10%
Mean proportion of missense SNVs (f_i)	0.0173	0.0254	0.0793	0.8780
Adjusted disruption rate	0.00350	0.00462	0.0132	0.0910

Supplementary Table C.2-2 Mean proportion of missense variants per individual across four MAF bins.

As noted earlier, most variants per individual genome, 87.8%, are very common (MAF > 10%). The *Adjusted disruption rate* per MAF bin listed in **Supplementary Table C.2-2** was obtained by multiplying the *Mean proportion of missense SNVs* by the *Disruption rate* listed in **Supplementary Table C.2-1**. Summing the *Adjusted disruption rate* across all MAF bins yields a *mean disruption rate per individual* = 11.2% ± 1.3%, where the error is calculated by the delta method.

C.2.2 Categorizing stable, moderately stable and unstable mutant proteins

Plate reader raw data from each 96-well plate consists of two fluorescence readings corresponding to GFP and mCherry expression in each well for proteins expressed in pDEST-DUAL vector. Wildtype/mutant groups are segregated to be on the same plate so they can be processed together. Each plate is allocated eight wells for background controls: four wells transfected with empty pDEST-DUAL vector such that only mCherry expression is expected, used as a GFP baseline, and four wells transfected with empty pcDNA-DEST47 vector where no expression is expected, used as mCherry baseline. All expression values are normalized as a z-score representing the number of standard deviations away from the mean background expression.

Supplementary Figure C.2-1 Representative plate reader scores. *All fluorescence readings are represented as a z-score away from the controls in that plate's 12th column. A12-D12 serve as GFP background using empty pDEST-DUAL vector, E12-H12 serve as mCherry background using empty pcDNA-DEST47 vector.*

Next, we apply basic quality control filters. A fluorescence reading is considered significant if the *P* value associated with its z-score on the background normal distribution is less than 0.05. We only perform analysis on experiments with significant wildtype expression for both GFP and mCherry channels. Further, we filter out any mutants that do not present significant mCherry expression.

We calculate wildtype activation and fold change to determine whether a mutant well is under-expressing GFP relative to its corresponding wildtype. Wildtype activation is the ratio between the GFP z-score and the mCherry z-score in the

wildtype well for that ORF and is reported as “wildtype stability score” in Figure 3.4-3A. Similarly, mutant activation is the ratio between the GFP z-score and mCherry z-score in the mutant well for an ORF and is reported as “mutant stability score” in Figure 3.4-3A. We then calculated fold change as the ratio of mutant activation over wildtype activation, reported throughout the text as the “stability score ratio.”

$$\text{WT Activation} = \frac{Z(\text{GFP}_{\text{WT}})}{Z(\text{mCherry}_{\text{WT}})}$$

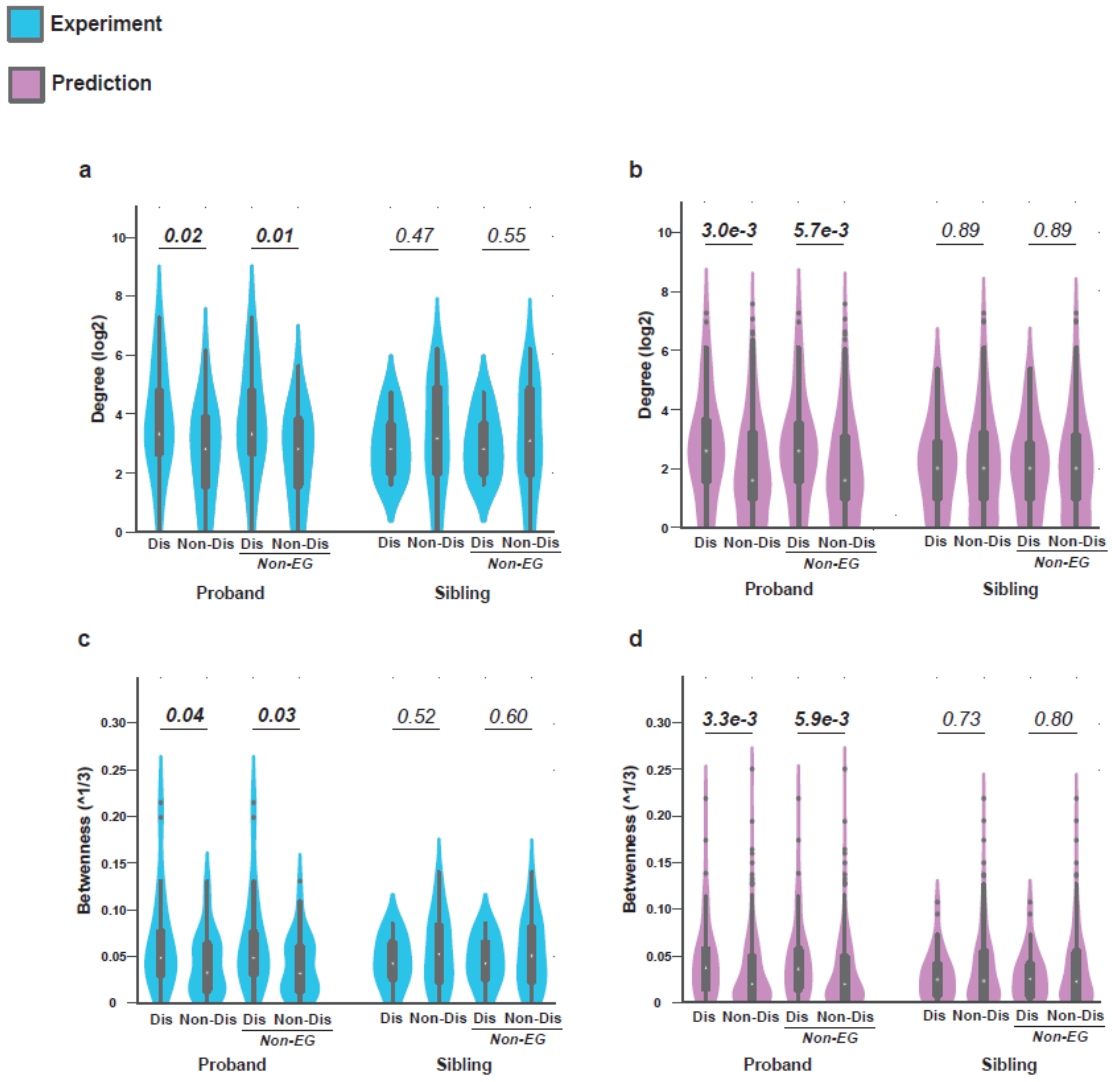
$$\text{Fold Change} = \frac{\frac{Z(\text{GFP}_{\text{mut}})}{Z(\text{mCherry}_{\text{mut}})}}{\frac{Z(\text{GFP}_{\text{WT}})}{Z(\text{mCherry}_{\text{WT}})}}$$

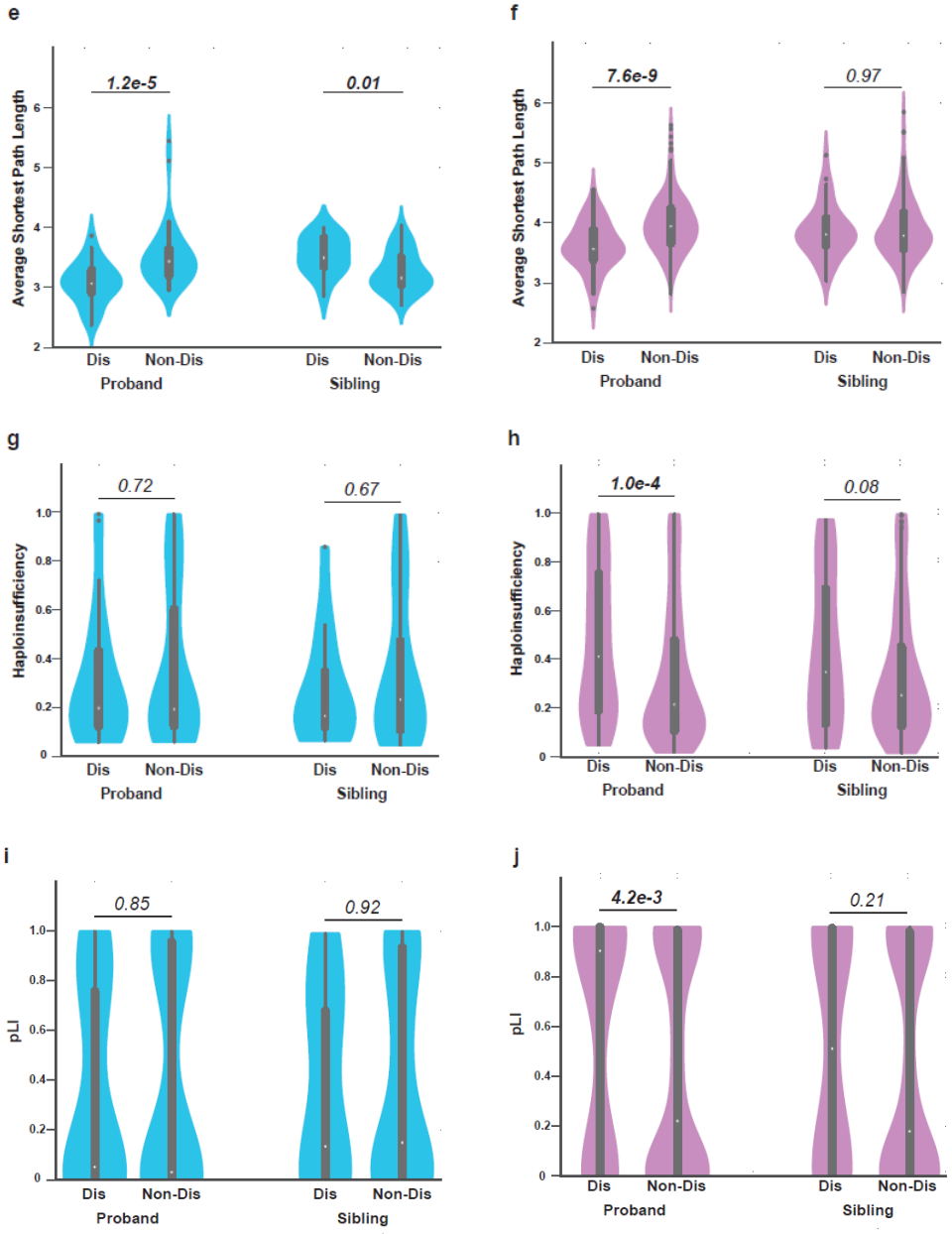
As an added quality control step, experiments with WT Activation less than 1.0 are removed. All other experiments are then classified into three groups: *stable* if the fold change is above 0.5, *moderately stable* if the fold change is between 0.5 and 0.0, and *unstable* if the fold change is less than 0.0.

APPENDIX D

SUPPORTING INFORMATION FOR CHAPTER 4

D.1 *Supplementary Figures for Chapter 4*





k

	Experiment	# of genes in SFARI	# of genes not in SFARI	OR	*P*
Proband	**Dis**	4	27	1.1	*0.57*
	Non-Dis	6	45		
Sibling	**Dis**	0	15	0.0	*1.0*
	Non-Dis	1	44		

l

	Prediction	# of genes in SFARI	# of genes not in SFARI	OR	*P*
Proband	**Dis**	17	56	3.2	***1.1e-3***
	Non-Dis	26	270		
Sibling	**Dis**	6	42	1.7	*0.23*
	Non-Dis	16	185		

m

Experiment

	Proband					Sibling				
	Dis (31)		Non-Dis (51)		*P*-value	Dis (15)		Non-Dis (45)		*P*-value
	Mean	s.e.m	Mean	s.e.m		Mean	s.e.m	Mean	s.e.m	
FMRP (794)	2.47	0.06	2.69	0.07	***0.025***	2.71	0.08	2.57	0.06	*0.91*
CHM (408)	2.40	0.06	2.62	0.07	***0.034***	2.65	0.08	2.48	0.06	*0.94*
EMB (1,865)	2.47	0.06	2.69	0.07	***0.025***	2.70	0.08	2.57	0.06	*0.90*
PSD (1,395)	2.47	0.06	2.68	0.07	***0.041***	2.68	0.08	2.56	0.06	*0.86*
SFARI (881)	2.53	0.06	2.75	0.07	***0.024***	2.78	0.08	2.62	0.06	*0.93*
SFARI hq (141)	2.46	0.06	2.67	0.07	***0.032***	2.72	0.08	2.54	0.06	*0.95*
DN65 (65)	2.53	0.06	2.76	0.07	***0.023***	2.79	0.08	2.62	0.06	*0.95*

n

Prediction

	Proband					Sibling				
	Dis (73)		Non-Dis (296)		*P*-value	Dis (48)		Non-Dis (201)		*P*-value
	Mean	s.e.m	Mean	s.e.m		Mean	s.e.m	Mean	s.e.m	
FMRP (794)	2.69	0.05	2.87	0.03	***1.4e-3***	2.80	0.06	2.82	0.03	*0.49*
CHM (408)	2.63	0.05	2.82	0.03	***1.6e-3***	2.73	0.06	2.77	0.04	*0.32*
EMB (1,865)	2.74	0.04	2.90	0.03	***4.0e-3***	2.83	0.06	2.86	0.04	*0.40*
PSD (1,395)	2.69	0.05	2.86	0.03	***2.2e-3***	2.80	0.06	2.82	0.03	*0.45*
SFARI (881)	2.77	0.05	2.94	0.03	***1.9e-3***	2.87	0.06	2.90	0.03	*0.42*
SFARI hq (141)	2.70	0.05	2.88	0.03	***1.2e-3***	2.80	0.06	2.82	0.04	*0.43*
DN65 (65)	2.79	0.05	2.96	0.03	***1.9e-3***	2.89	0.06	2.91	0.03	*0.39*

Supplementary Figure D.1-1 Analyses of interaction-disrupting (Dis) and non-disrupting (Non-Dis) dnMis mutations measured experimentally (blue) or computationally (purple).

(a-b) Degree and (c-d) betweenness distributions of proteins with Dis or Non-Dis dnMis mutations in ASD probands and unaffected siblings. Non-EG: non-essential gene-encoded proteins. P-values were calculated using two-tail U-test (P < 0.05 in ***bold****). (e-f) Average shortest path length distributions of proteins with Dis or Non-Dis dnMis mutations in ASD probands and unaffected siblings. P-values were calculated using two-tail U-test (P < 0.05 in* ***bold****). (g-h) Haploinsufficiency and (i-j) pLI of genes that harbor Dis or Non-Dis dnMis mutations in ASD probands and unaffected siblings. P-values were calculated using two-tail U-test (P < 0.05 in* ***bold****). (k-l) Contingency tables for the counts of genes harboring Dis or Non-Dis dnMis mutations in SFARI database. P-values were calculated using one-tail Fisher's exact test (P < 0.05 in* ***bold****, OR: Odds Ratio). (m-n) Distance of proteins with Dis or Non-Dis dnMis mutations to seven classes in a protein interactome network background. Number of proteins in each class is indicated in parentheses. P-values were calculated using one-tail U-test (P < 0.05 in* ***bold****).*