UNDERSTANDING MECHANISMS OF HUMAN DISEASES THROUGH BIOLOGICAL NETWORKS

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

Yu Guo

August 2015

UNDERSTANDING MECHANISMS OF HUMAN DISEASES THROUGH BIOLOGICAL NETWORKS

Yu Guo, Ph. D.

Cornell University 2015

In the past few decades, great progress has been made in uncovering the molecular bases of many human diseases. As we begin to appreciate the complex cellular architecture, the "one-gene/one-enzyme/one-function" concept is no longer apt to explain the complex genotype-phenotype relationships. In fact, the cell functions as an intricate network of interacting genes, gene products and metabolites. Perturbations in biological networks may underlie many disease phenotypes. My dissertation examines the mechanisms of human diseases and disease mutations in the context of biological networks.

First, I investigated the regulatory effects of transcription factors (TFs) and microRNAs (miRNAs) on target gene expression, protein-protein interaction and disease association in an integrated gene regulation network. My results suggest that TFs and miRNAs occupy distinct niches in the overall regulatory network within the cell. While TFs tend to regulate intra-module clusters, miRNAs tend to regulate inter-module clusters.

Next, to better understand different molecular mechanisms through which mutations lead to diseases, I examined the effects of disease-associated mutations with different inheritance modes and molecular types in a three-dimensional protein

interactome network. I found that although recessive mutations on the interaction interface of two interacting proteins tend to cause the same disease, this widely-accepted "guilt-by-association" principle does not apply to dominant mutations. Furthermore, my analyses suggest that a significant fraction of truncating mutations, which are often considered as "loss-of-function" mutations, can generate functional protein products.

Then, I investigated the molecular phenotypes of human disease mutations that are native in the orthologous proteins of other species. As these mutations are potentially compensated in the other species through epistatic selection, they are named potentially epistatic mutations (PEMs). Here, we experimentally demonstrated that PEMs are less deleterious than regular disease mutations, and identified potential intra- and inter-protein compensatory mutations for the PEMs.

Finally, I set up a variant prioritizing pipeline that incorporates various biological data such as known disease genes, biological pathways, protein-protein interactions and protein structures to identify predisposing mutations from the whole-genome and whole-exome sequences of Crohn disease and multiple myeloma patients. This analysis pipeline led to the identification of a novel, high-risk variant in Crohn disease.

BIOGRAPHICAL SKETCH

I was born in Inner Mongolia, China in 1986, and moved to Singapore with my family in 1998. I attended the Nanyang Technological University from 2006 to 2010, where I graduated with a first class honors B.S. degree in Biological Sciences. I conducted my honors thesis research in the Gurdon Institute at Cambridge University, where I studied the primitive endoderm formation in early mouse embryos. In 2010, I started graduate studies in Cornell University under the supervision of Dr. Haiyuan Yu, and spent the next five years studying the molecular mechanisms of human disease mutations using systems biological approaches in the context of biological networks.

To my mother Lixia and my husband Ernest

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor, Dr. Haiyuan Yu for his guidance and encouragement for the past five years. His passion for science has been a great inspiration. I am grateful to my co-advisor Dr. Andrew Clark and my special committee members Drs. Andrew Grimson and Jason Mezey for providing great advice on my thesis projects, and valuable suggestions that improved my manuscripts. In addition, I thank my wonderful collaborators Drs. Steven Lipkin and Xiaomu Wei for many fruitful discussions. My thesis would not be possible without the help and support from the past and current members of the Yu lab. I thank them for creating a vibrant, intellectual and supportive environment in the lab, and for their friendship.

I am deeply grateful to my family for their encouragement and support. I thank my grandfather for inspiring me and encouraging me to pursue my interests, and my uncle Hongwu and family for their support during my PhD studies. I am especially grateful to my mother, who has been my mentor, friend and confidant. Lastly, I thank my loving husband Ernest who spent the past five years with me at Cornell. I could not be more grateful for his loving support and companionship and I am glad that we experienced graduate school together and made it a blissful five years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLE

CHAPTER1

INTEGRATED NETWORK ANALYSIS REVEALS DISTINCT REGULATORY ROLES OF TRANSCRIPTION FACTORS AND MICRORNAS

Guo Y, Alexander K, Clark AG, Grimson A and Yu H. In Revision.

1.1. Abstract

Analysis of transcription regulatory networks has revealed many principal features that govern gene expression regulation. MicroRNAs (miRNAs) have emerged as another major class of gene regulators that influence gene expression post-transcriptionally, but there remains a need to assess quantitatively their global roles in gene regulation. Here, we have constructed an integrated gene regulatory network comprised of transcription factors (TFs), miRNAs and their target genes, and analyzed the effect of regulation on target gene expression, protein-protein interaction and disease association. We found that while target genes regulated by the same TFs tend to be co-expressed, co-regulation by miRNAs does not lead to co-expression. Analysis on interacting protein pairs in the regulatory network revealed that compared to genes co-regulated by miRNAs, a higher fraction of genes co-regulated by TFs encode proteins in the same complex. Although these results suggest that genes co-regulated by TFs are more functionally related than those co-regulated by miRNAs, genes that share either TF or miRNA regulators are more likely to cause the same disease. Further analysis on the interplay between TFs and miRNAs suggest that TFs tend to regulate intra-module/pathway clusters, while miRNAs tend to regulate inter-module/pathway clusters. These results demonstrate that although TFs and miRNAs both regulate gene expression, they occupy distinct niches in the overall regulatory network within the cell.

1.2. Introduction

Gene expression is controlled and fine-tuned at multiple levels in a hierarchical gene

regulatory network. Transcription factors (TF) activate or repress gene expression by binding transcription factor binding sites (TFBS) in gene promoters or *cis*-regulatory modules [1,2]. TFs were believed to be the primary regulators of gene expression until research in the past decade revealed miRNA as another major class of gene expression regulator [3]. miRNAs are small, non-coding RNAs that fine-tune gene expression post-transcriptionally. Mature miRNAs bind complementary sequences of target mRNAs, causing mRNA degradation and/or translation repression [3]. miRNAs regulate many biological processes and have been implicated in the development of human diseases including cancer [3-5]. Recent research suggested that the majority of human genes might be targets of miRNAs [6]. Many miRNA targets are TFs, which can in turn regulate miRNA expression, forming an intricate regulatory network.

As the canonical gene expression regulator, TFs have been well characterized, and system-wide properties of the transcription regulatory network have been explored [7-9]. In recent years, researchers are increasingly interested in the combinatorial interactions between TFs and miRNAs. An integrated regulatory network that includes both transcriptional and post-transcriptional regulation is necessary to provide a more complete picture of gene expression regulation and may reveal basic regulatory principles underlying disease phenotypes. Recent studies have found recurring co-regulatory motifs involving both TFs and miRNAs, such as TF-miRNA co-regulating pairs and feed-forward loops, indicating prevalent crosstalk and cooperation between these two modes of gene regulation [10-12].

As genome-wide transcription-factor-binding data were not readily available, previous studies on the integrated regulatory network inferred TF-gene regulatory

relationships from computationally predicted TFBS. Recently, large-scale ChIP-seq experiments from the ENCODE project generated system-wide data on transcription factor binding patterns [13]. Making use of the transcription factor binding data from the ENCODE project, Gerstein et al. (2012) created a human transcriptional regulatory network [14]. By combining this ENCODE transcriptional regulation network and high confidence miRNA target predictions, we constructed a human integrated regulatory network to investigate the possible differences in the roles of TFs and miRNAs in gene regulation and the synergistic actions of these two types of regulators. Previous studies of TF-miRNA co-regulation in the regulatory network suggested relationships between regulation and gene expression [10-12], but downstream effects of regulation, especially at the organismal level, remain unclear. Here we studied the effects of TF and miRNA regulation at three levels: gene expression, protein-protein interaction and organism-level disease phenotypes, and we found that TFs and miRNAs exhibit distinct roles in the regulation of gene expression.

1.3. Results

1.3.1. Integrated gene regulatory network

The TF-gene regulatory relationships were derived from ENCODE data generated by ChIP-Seq experiments. The high confidence set of TF-gene regulatory relationships were downloaded from the supplementary website of Gerstein et al. (2012)[14]. Among the 119 transcription-related factors studied in the ENCODE project, we only considered sequence specific transcription factors that recognize and bind to specific DNA sequence motifs.

Currently, several major miRNA target prediction algorithms, such as TargetScan [15], miRanda [16], PicTar [17] and PITA [18], make genome-wide predictions of miRNA targets based on target site conservation. Studies of protein levels after miRNA knockdown and transfection demonstrated that TargetScan outperforms other miRNA prediction algorithms in terms of prediction accuracy [19,20]. In this study, we obtained miRNA identities and their predicted targets from TargetScan. To ensure only high confidence miRNA-gene regulatory relationships are included, we filtered TargetScan predictions based on both conservation criteria [6,15] and estimates of site performance, referred to as Context Score [21]. We used only gene targets with a Pct $\geq$ 0.5, which indicates that there is at least a 50% probability that a sequence is selectively maintained as a miRNA target site, and a Context Score $\leq$ -0.2. Grimson et al. (2007) demonstrated by siRNA transfection experiments that predicted miRNA targets with lower Context Score are more down-regulated in response to siRNA expression. Context Scores of -0.2 or lower were chosen as the cutoff because such targets were measurably down regulated (about 25% change in expression on average for conserved sites) in the siRNA transfection experiments [21]. To avoid redundancies in the network and subsequent analyses, mature miRNAs with identical seed regions were grouped into miRNA families based on the miRNA family information from TargetScan.

Combining the two types of regulatory relationships, we constructed a human integrated gene regulatory network with a total of 35,304 regulatory relationships among 83 TFs, 77 miRNA families and 11,407 target genes.

1.3.2. Synergistic effects of TF and miRNA regulation on target gene expression, protein-protein interaction and disease association

Previous studies have established that both TFs and miRNAs work cooperatively to regulate their gene targets [22,23]. Here we investigated the additive effects of TF regulation and miRNA regulation on their target genes.

First, we examined the expression relationships among genes regulated by the same TF(s) or miRNA(s). Previous studies have shown in multiple organisms that genes regulated by the same TF tend to be co-expressed [24-27]. Here, we found that the degree of expression correlation depends on the number of shared regulators. We calculated the log odds ratio of co-expression of gene pairs co-regulated by one or more TFs compared to random gene pairs. We found that the likelihood of two genes being co-expressed increases with the number of common TFs that regulate them (Figure 1.1A). On the other hand, genes regulated by the same miRNAs do not tend to be co-expressed compared to random gene pairs (Figure 1.1B). Even gene pairs co-regulated by four or more miRNAs are not more likely to co-express (LOD = -0.04, P=0.45). This observation is the opposite of the commonly accepted view that co-regulation leads to co-expression. However, the observation is consistent with our current understanding of miRNAs: they only have moderate repressive effect on target gene transcript levels and do not control the on/off state of the target gene transcription [3]. Furthermore, the overall effect of miRNA regulation of gene expression is much more subtle compared to TF regulation [3].

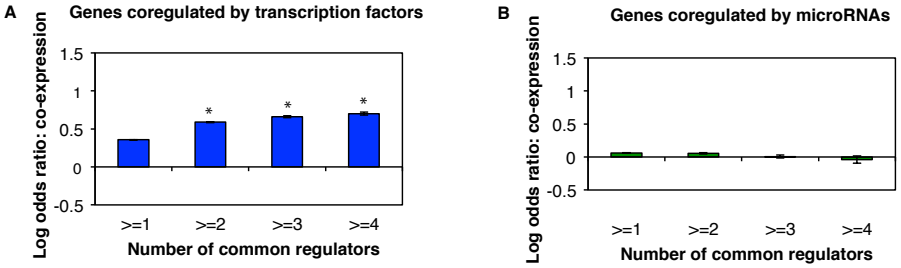


Figure 1.1 Synergistic effects of TFs and miRNAs on target gene co-expression. LOD values are calculated for the enrichment of co-expression relationships between gene pairs co-regulated by multiple (**A**) TFs or (**B**) miRNAs. **P* < 0.05. *P*-values are calculated by the *Z*-test. Error bars indicate ±SE.

The level of gene expression affects the amount of protein translated in the cell [28]. Proteins are the functional units of the cell and carry out their functions through interactions with other proteins. To investigate the effects of regulation on protein function, we studied the interactions between protein products of genes regulated by the same TF(s) or miRNA(s). We found that the protein products of genes co-regulated by either TFs or miRNAs are significantly more likely to physically interact with each other compared to random expectation, and the likelihood of physical interaction increases with the number of shared regulators (Figure 1.2A-B). For example, protein products of genes co-regulated by three or more common regulators are almost twice as likely to physically interact compared to random expectation (OD=1.88, $P < 10^{-6}$ for TF regulation; OD=1.99, $P < 10^{-4}$ for miRNA regulation). Furthermore, protein products of genes co-regulated by more TFs or miRNAs tend to be in closer proximity in the protein interaction network (Figure A.1). Within the cell, proteins can form stable complexes or dynamically interact with each other to carry out subcellular functions. The stable interactions bring proteins into tightly regulated

functional modules, while transient interactions connect and coordinate these modules [29]. To investigate the relationship between regulation and protein interaction dynamics, we identified stable complexes in the protein-protein interaction network using the ClusterONE algorithm [30]. We found that genes co-regulated by the same miRNAs are less likely to encode proteins in the same complex compared to genes co-regulated by TFs (Figure 1.2C, $P < 10^{-5}$). Previous studies found that genes encoding subunits in the same protein complex are globally co-expressed [31,32]. On the other hand, genes encoding proteins involved in transient interactions only co-express under specific conditions and do not have highly correlated expression profiles [29]. This explains our observation that although genes co-regulated by the same miRNAs are more likely to encode for interacting proteins, co-regulation by miRNAs has little effect on the global co-expression of target genes.

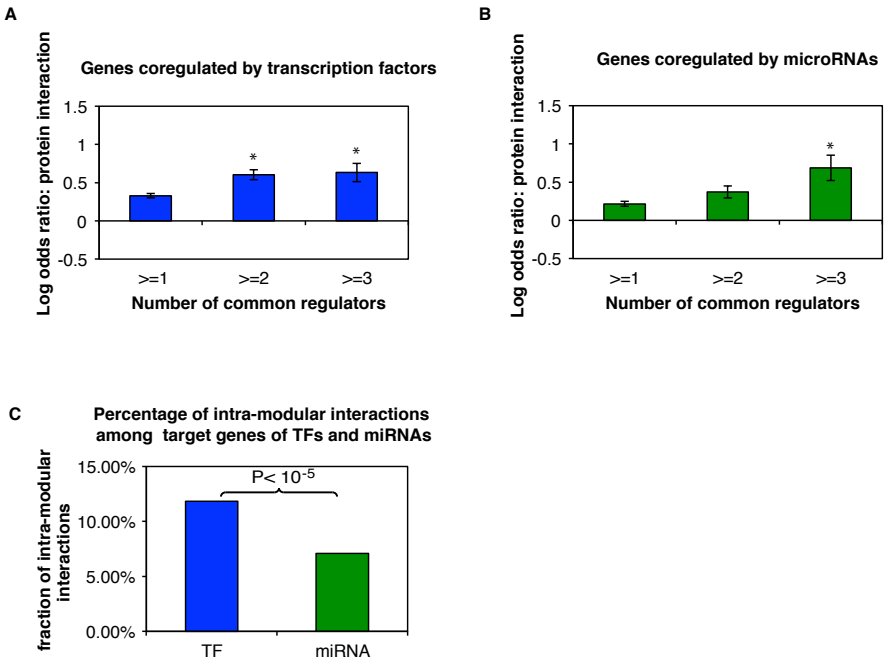


Figure 1.2 Synergistic effects of TFs and miRNAs on the protein-protein interaction of targets. LOD values are calculated for the enrichment of protein-protein interactions

between gene pairs co-regulated by multiple TFs through (**A**) TFs or (**B**) miRNAs. **P* < 0.05. *P*-values are calculated by the *Z*-test. Error bars indicate ±SE. (**C**) The percentage of intra-modular interactions among interacting protein pairs encoded by co-regulated genes.

Mis-regulation of gene expression often leads to disease phenotypes. Both TFs and miRNAs have been found to be associated with a wide range of human diseases, including the development and progression of cancers [33,34]. However, the relationship between regulatory network architecture and the involvement of genes in different human diseases is not well studied at a genome-wide scale. To this end, we compiled a comprehensive list of disease genes and their associated diseases from HGMD and OMIM databases, and studied the relationship between co-regulation and co-association of genes to diseases. Specifically, we calculated the enrichment of co-regulated disease genes that are associated with the same disease with respect to random expectations. Genes jointly regulated by more TFs are more likely to be associated with the same diseases (LOD= 0.32, 0.75, 0.96, and 1.25 for gene pairs jointly regulated by more than 1, 2, 3, and 4 TFs respectively; Figure 1.3A). Similarly, genes co-regulated by multiple miRNAs are also more likely to cause the same diseases (LOD= 0.22, 0.57, and 1.24 for genes jointly regulated by more than 1, 2, and 3 miRNAs respectively; Figure 1.3B).). Together with the expression and interaction analyses above, our results demonstrate that genes co-regulated by multiple TFs tend to be more related on all three functional levels we examined. This shows that genes sharing more regulating TFs tend to be more functionally similar, and tend to form tightly regulated modules that function together in the same biological processes/pathways. In contrast, miRNAs do not regulate the coordinated expression

of genes in the same functional module and tend to regulate inter-complex protein-protein interactions, but target genes co-regulated by the same miRNAs are still significantly more likely to be associated with the same disease. This suggests that miRNAs play an important role in inter-modular regulation, where they coordinate target genes in related biological processes/pathways. Overall, our results reveal significant synergistic effects of both TF and miRNA regulation, and highlight the importance of combinatorial regulations by TFs and miRNAs in biological processes.

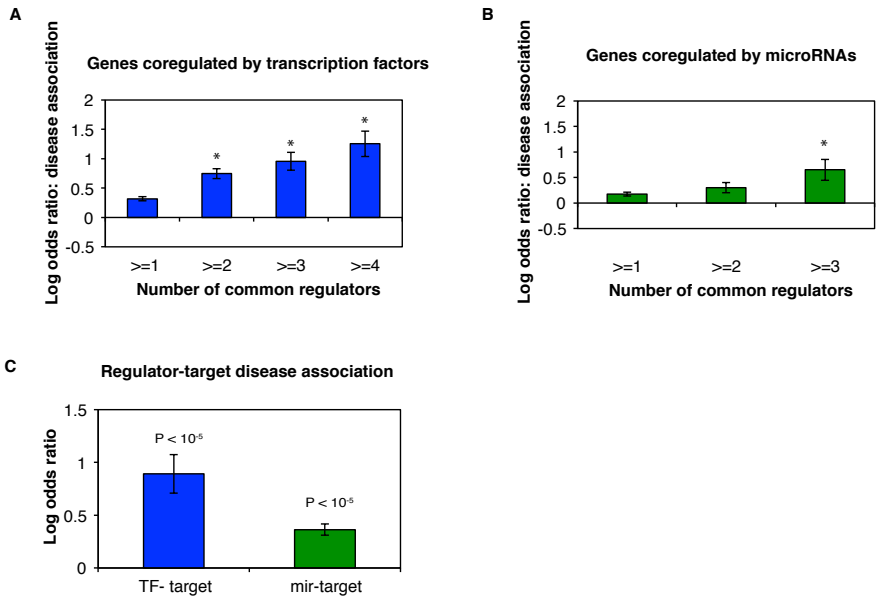


Figure 1.3 Effects of TF- and miRNA- regulation on target gene disease association. LOD values of the enrichment for co-regulated genes to cause the same disease are calculated for gene pairs co-regulated by multiple TFs through (**A**) TFs or (**B**) miRNAs. **(C)** LOD values of the enrichment for regulator-gene pairs to be associated with the same disease compared to random expectations. *$P < 0.05$. *P*-values are calculated by the *Z*-test. Error bars indicate ±SE.

To verify the statistical significance of our observations, we performed randomization tests to evaluate the significance of the enrichments (see Methods), and obtained results consistent with those described above (Figure A.2). Although

TargetScan provides a genome-wide, unbiased prediction of miRNA targets, most of the targets predicted are not verified experimentally. To further validate our results, we generated a high-quality set of experimentally verified miRNA-gene interactions using manually curated data from Tarbase [35] and miRTarBase [36]. To avoid study bias arising from small-scale, gene-specific experiments, we only included data generated by high-throughput experimental methods such as CLIP-seq and Degradome-seq. We repeated all our calculations with the high-quality experimentally verified miRNA target set and found the results to be consistent with those described above (Figure A3).

1.3.3. Functional relationship between regulators and their target genes

Genes associated with the same disease tend to have correlated gene expression and their protein products tend to physically interact, forming functional modules in the cellular network [37,38]. As TFs and miRNAs control the timing and level of expression of their target genes, we postulate that disruption of the regulator functions and disruption of the target protein functions are likely to result in the same diseased state. By comparing the diseases associated with TFs and the diseases associated with their gene targets, we found that overall, TFs are significantly more likely to cause the same disease as their gene targets compared to random TF-target pairs (LOD= 0.89, $P < 10^{-5}$; Figure 1.3C). To perform the same analysis on miRNAs, we compiled a list of literature curated miRNA-disease associations from the HMDD [39] and miR2Disease [40] databases. Similarly, we found that miRNAs are also significantly more likely to be associated with the same disease as the genes they regulate (LOD= 0.36, $P < 10^{-10}$;

Figure 1.3C). This confirms our hypothesis that disruptions of TF or miRNA functions tend to have similar effects as disruptions of their target gene functions.

1.3.4. Different roles of TF and miRNA regulation: Intra-modular vs. Inter-modular

Recent studies found that crosstalk and cooperation between TFs and miRNAs are highly prevalent and could be an integral part of the gene regulatory network [10,41-43]. To further understand the crosstalk between TFs and miRNAs, we investigated the regulatory relationships between TFs and miRNAs. From our gene regulatory network, we found 1004 miRNA pairs co-regulated by at least one common TF, and 262 TF pairs co-regulated by at least one common miRNA.

First we examined the functional similarity of TFs that are targeted by the same miRNA(s). We computed the fraction of shared targets for each TF pair (number of shared targets / total number of targets of the two TFs), and compared the distribution of the fraction of shared targets of TF pairs regulated by the same miRNA and that of random TF pairs. We found that TF pairs regulated by the same miRNA are not more likely to share targets compared to random TF pairs (P = 0.86 by Wilcoxon rank sum test; Figure 1.4A), suggesting that TFs regulated by the same miRNA may regulate different biological processes. However, TFs do tend to share targets with the miRNA that regulates them, implying functional overlap between a miRNA and the TFs it regulates (Figure A.4). In contrast, about 81% of miRNA pairs regulated by the same TF share gene targets, which is significantly higher than random expectation ($P < 10^{-8}$; Figure 1.4C). This shows that miRNAs regulated by the same

TFs tend to have higher functional overlap. As a comparison, we also calculated the distribution of the fraction of shared targets of TF pairs regulated by the same TF. We found that TFs that are regulated by the same TF tend to share more gene targets compared to random TF pairs ($P < 10^{-4}$; Figure 1.4B), suggesting that TFs regulated by the same TF also tend to have related functions. In summary, TF and miRNA pairs that are co-regulated by an upstream TF tend to target the same genes, whereas TF pairs co-regulated by the same miRNAs do not tend to share more targets compared to random TF pairs.

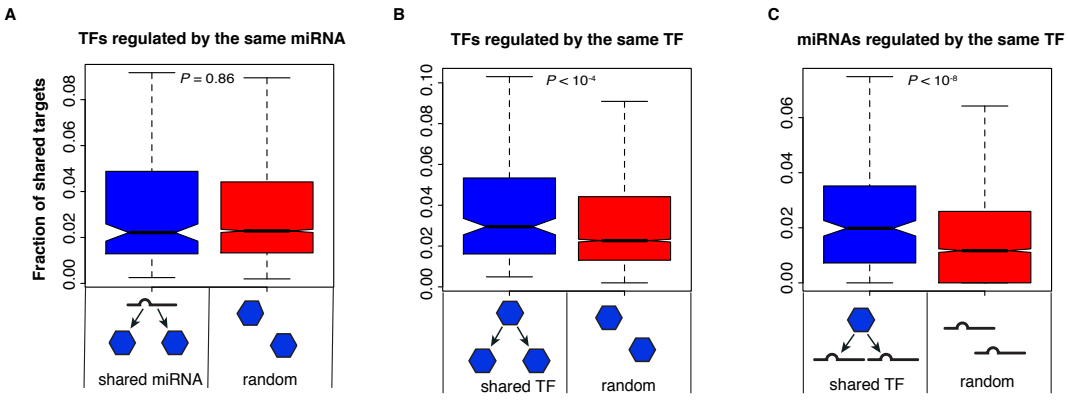


Figure 1.4 Inter-regulation of TFs and miRNAs. Distribution of the fraction of shared targets of **(A)** TF pairs regulated by the same miRNA, **(B)** TF pairs regulated by the same TF, and **(C)** miRNA pairs regulated by the same TF. *P*-values are calculated using the Wilcoxon rank sum test.

To further verify the above observations, we formulated a functional similarity metric for TFs and miRNAs based on the overlap of their target gene sets, the proximity of their target genes in the protein-protein interaction network and the similarity of their target genes according to the Gene Ontology biological processes terms (See methods). Consistent with our previous observations, we found that TFs

regulated by the same miRNAs are not more functionally similar than random TF pairs, while miRNA pairs ($P < 10^{-15}$) and TF pairs ($P < 10^{-12}$) regulated by the same TF are significantly more similar than random miRNA pairs or TF pairs without regulation (Wilcoxon rank sum test; Figure A.5). This suggests that TFs regulate groups of functionally similar downstream regulators that tend target the same genes or functionally related genes. In contrast, miRNAs regulate functionally disparate TFs that are likely to be involved in different biological processes.

Thus far, we found that genes co-regulated by TFs form tightly regulated functional modules, and that miRNAs/TFs regulated by the same TF are functionally similar. These results suggest that TFs tend to regulate genes within the same functional module, where genes regulated by the same TF participate in the same biological process/pathway (Figure 1.5A). Examples of intra-modular multi-level regulation is shown in Figure 1.5A. The oncogenic transcription factor *MYC* is involved in many different types of human cancers [44]. For example, overexpression of *MYC* is associated with human prostate cancer [45,46]. It has been found that *MYC* regulates the expression of many tumor-suppressing miRNAs in lymphoma and prostate cancer cells [45,47]. In our integrated regulatory network, *MYC* regulates *hsa-miR-19b* and *hsa-miR-92a*, which are themselves also associated with prostate cancer. These miRNAs in turn regulate *PTEN*, a tumor suppressor gene that was found to be inactivated in somatic prostate cancers [48]. Here, the transcription factor *MYC* regulates a group of miRNAs, which in turn regulate the downstream gene *PTEN* in the same disease module. In this disease module, we also found another miRNA, *hsa-miR-19a*, that has not been previously associated with prostate cancer. As *hsa-miR-19a* is in the

same miRNA family as *hsa-miR-19b*, it is likely that it also plays a role in prostate cancer.

On the other hand, genes regulated by the same miRNA(s) do not tend to co-express globally, their protein products tend to interact inter-modularly, but they are still more likely to be associated with the same disease. In addition, a single miRNA may regulate multiple TFs that carry out different functions. Taken together, our results suggest that miRNAs are involved in inter-modular regulation, where genes regulated by the same miRNA may not necessarily be in the same protein complex or pathway but are involved in related cellular processes (Figure 1.5B). An example of inter-modular multi-level regulation is shown in Figure 1.5B.

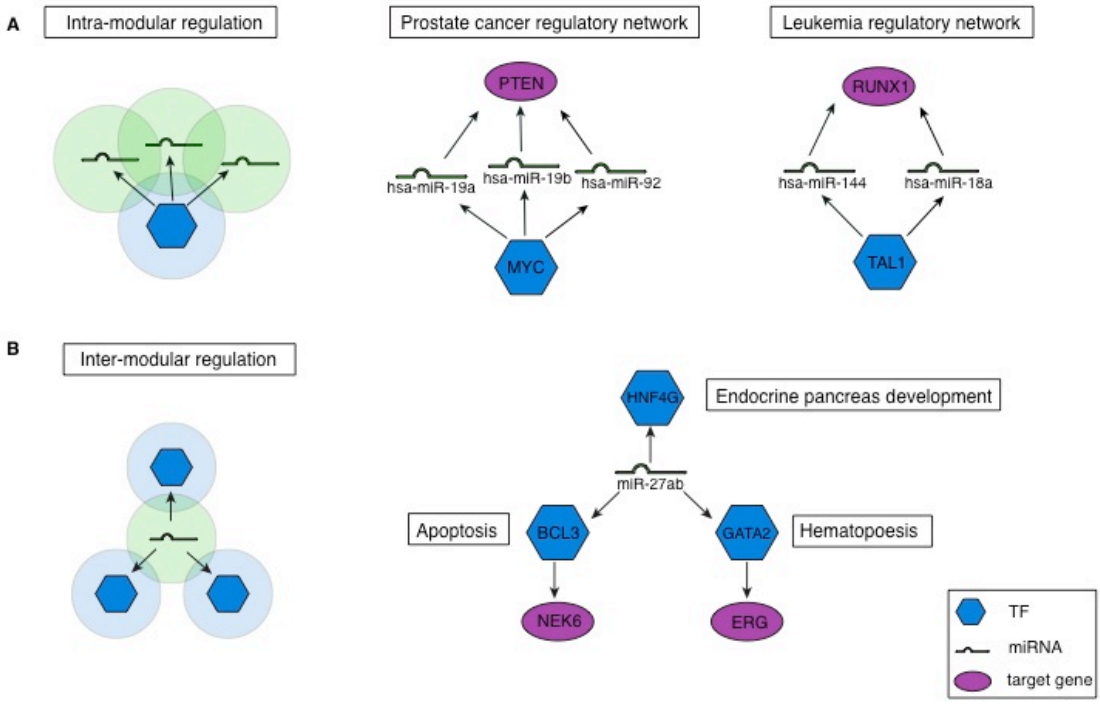


Figure 1.5 Intra-modular regulation and inter-modular regulation **(A)** Schematic and an example of intra-modular regulation by TF. Transcription factor *MYC* and downstream targets are associated with prostate cancer. **(B)** Schematic and an example of inter-modular regulation by miRNA. *miR-27ab* targets three different TFs

that each regulates different biological processes.

1.4. Discussion

In this study, we have constructed an integrated gene regulatory network comprising both transcriptional regulation and post-transcriptional regulation, and investigated on a global scale, the differences between the two layers of regulation on three functional levels. Our results revealed that TFs are involved in intra-modular regulation, where multiple TFs act cooperatively to regulate a set of genes that tend to co-express, interact physically and associate with the same diseases. On the other hand, miRNAs coordinate related cellular pathways/processes through inter-modular regulation. Gene targets regulated by the same miRNA(s) show higher expression variability, they tend not to encode proteins in the same complex, although their protein products are in closer proximity in the protein-protein interactome network; and they are more likely to be associated with the same diseases. A previous study by Liang and Li on the correlation between miRNA regulation and protein-protein interactions found that miRNAs have a higher propensity to target inter-modular protein hubs compared to intra-modular protein hubs [49], which further supports our observations. miRNA regulation of genes across different functional modules/pathways is biologically important, as different functional modules have diverse expression profiles and regulation is essential for the coordination among different functional modules in the cell. Indeed, clustering of miRNAs by functional similarity revealed that disease-associated miRNAs tend to be at the interface between adjacent functional modules compared to non-disease-associated miRNAs [50].

In conclusion, we found that although TFs and miRNAs share similar regulatory logic, such as synergistic regulation in network motifs, they appear to occupy distinct niches in the gene regulatory network, and that these differences impact the role that TFs and miRNAs play in mediating disease risk. Our findings provide new insights into the global architecture and organization principles of the gene regulatory network.

1.5. Materials and Methods

1.5.1. TF and miRNA regulation datasets

We obtained regulatory relationships between 77 conserved miRNA families and 5858 predicted targets from TargetScan, a leading miRNA target prediction algorithm, based on both conservation criteria [6,15] and estimates of site performance, referred to as Context Score [21]. The high quality, an experimentally verified miRNA target set was obtained from Tarbase[35] and miRTarBase[36]. In total, there are 6,470 interactions among 59 miRNA families and 3660 target genes that are verified by high throughput experiments.

The transcriptional regulatory network are derived from ChIP-seq data generated by the ENCODE project. The TF regulatory networks were downloaded from the supplementary website of Gerstein et al. (2012)[14]. To ensure the quality of the network, we used the filtered set of high confidence TF-gene associations. The transcriptional regulatory network comprises 83 sequence specific transcription factors and 8243 target genes.

1.5.2. Gene expression data and PCC calculation

We used the gene expression data generated by the Genomics Institute of Novartis Research Foundation (GNF) GeneAtlas project, which measured gene expression profiles of 79 human tissues. The gene expression data is downloaded from the Gene Expression Omnibus (GDS596). We quantile-normalized gene expression values, and took the average value of the probes for each gene. Pearson correlation coefficient (PCC) was calculated for all possible pairs of genes using a massively parallel Java program [29]. Two genes are considered to be co-expressed if they have PCC of 0.3 or greater (top 6% of all gene pairs).

1.5.3. Protein-protein interaction and disease-gene association data

High-quality, binary protein-protein interactions were obtained from HINT [51], a protein interaction database with high-quality interactions collated from literature-curation and high-throughput experiments.

A comprehensive list of disease-associated genes was compiled from the Human Gene Mutation Database (HGMD) [52,53] and the Online Mendelian Inheritance in Man (OMIM) [54,55] database. miRNA-disease associations were collated from two databases: The human microRNA disease database (HMDD) [39] and miR2Disease database [40]. In total we collected 2,712 manually curated associations between 263 miRNAs and 184 diseases. To standardize the disease nomenclature across databases, unique disease identifiers were assigned to each phenotypically distinct disease through computational and manual curation.

1.5.4. Calculating the similarity score of two regulators

The similarity score of two regulators were calculated based on: 1) the number of gene targets they share, 2) the number of targets of each regulator that interacts with target proteins of the other regulator in the protein-protein interaction network, and 3) the number of targets of each regulator that are functionally similar to the targets of the other regulator based on biological process terms from Gene Ontology. The similarity score of two regulators is computed by dividing the total number of target genes of the two regulators that satisfies 1, 2 or 3 of the criteria described above, by the total number of number of target genes of the two regulators.

Functional similarity of two genes based on Gene Ontology biological process was calculated using the total ancestry measure as previously described [29,56]. Functional similarities between all genes were calculated using a massively parallel Java program [29].

1.5.5. Statistical analyses

The enrichment of co-expression, protein-protein interaction and disease association among co-regulated genes with respect to random expectations were measured by log odds ratio (LOD).

$$LOD = \ln\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right)$$

Where p_1 is the fraction of co-regulated gene pairs sharing a specific functional relationship (i.e. co-expression, protein-protein interaction or association with the same diseases), and p_2 is the fraction of all possible gene pairs sharing the functional

relationship. The statistical significance of the enrichment was evaluated by the *Z*-test.

To verify the correctness of our statistical model, we also performed randomization tests to evaluate the enrichment of co-expression, protein-protein interaction and disease association among co-regulated genes. We generated 100 random networks keeping the degree distribution and network topology intact. For each functional relationship, we compared the fraction of co-regulated gene pairs sharing a specific functional relationship in the real network to the average fraction of co-regulated gene pairs sharing the same functional relationship in random networks. We found that using random networks as control yields the same results as using all possible gene pairs as controls as described above (Figure A.2).

REFERENCES

1 Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annual review of genomics and human genetics* **7**, 29-59, doi:10.1146/annurev.genom.7.080505.115623 (2006).

2 Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99-104, doi:10.1038/nature02800 (2004).

3 Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281-297 (2004).

4 Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B. & Cohen, S. M. bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell* **113**, 25-36 (2003).

5 Flynt, A. S. & Lai, E. C. Biological principles of microRNA-mediated regulation: shared themes amid diversity. *Nature reviews. Genetics* **9**, 831-842, doi:10.1038/nrg2455 (2008).

6 Friedman, R. C., Farh, K. K., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome research* **19**, 92-105, doi:10.1101/gr.082701.108 (2009).

7 Lee, T. I. *et al.* Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804, doi:10.1126/science.1075090 (2002).

8 Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics* **31**, 64-68, doi:10.1038/ng881 (2002).

9 Yu, H. & Gerstein, M. Genomic analysis of the hierarchical structure of regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 14724-14731, doi:10.1073/pnas.0508637103 (2006).

10 Shalgi, R., Lieber, D., Oren, M. & Pilpel, Y. Global and local architecture of the mammalian microRNA-transcription factor regulatory network. *PLoS computational biology* **3**, e131, doi:10.1371/journal.pcbi.0030131 (2007).

11 Su, N., Wang, Y., Qian, M. & Deng, M. Combinatorial regulation of transcription factors and microRNAs. *BMC systems biology* **4**, 150, doi:10.1186/1752-0509-4-150 (2010).

12 Tsang, J., Zhu, J. & van Oudenaarden, A. MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Molecular cell* **26**, 753-767, doi:10.1016/j.molcel.2007.05.018 (2007).

13 ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

14 Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100, doi:10.1038/nature11245 (2012).

15 Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15-20, doi:10.1016/j.cell.2004.12.035 (2005).

16 John, B. *et al.* Human MicroRNA targets. *PLoS biology* **2**, e363, doi:10.1371/journal.pbio.0020363 (2004).

17 Krek, A. *et al.* Combinatorial microRNA target predictions. *Nature genetics* **37**, 495-500, doi:10.1038/ng1536 (2005).
18 Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. & Segal, E. The role of site accessibility in microRNA target recognition. *Nature genetics* **39**, 1278-1284, doi:10.1038/ng2135 (2007).
19 Selbach, M. *et al.* Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58-63, doi:10.1038/nature07228 (2008).
20 Baek, D. *et al.* The impact of microRNAs on protein output. *Nature* **455**, 64-71, doi:10.1038/nature07242 (2008).
21 Grimson, A. *et al.* MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell* **27**, 91-105, doi:10.1016/j.molcel.2007.06.017 (2007).
22 Hobert, O. Gene regulation by transcription factors and microRNAs. *Science* **319**, 1785-1786, doi:10.1126/science.1151651 (2008).
23 Martinez, N. J. & Walhout, A. J. The interplay between transcription factors and microRNAs in genome-scale regulatory networks. *BioEssays : news and reviews in molecular, cellular and developmental biology* **31**, 435-445, doi:10.1002/bies.200800212 (2009).
24 Yu, H., Luscombe, N. M., Qian, J. & Gerstein, M. Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends in genetics : TIG* **19**, 422-427, doi:10.1016/S0168-9525(03)00175-6 (2003).
25 Gu, Q., Nagaraj, S. H., Hudson, N. J., Dalrymple, B. P. & Reverter, A. Genome-wide patterns of promoter sharing and co-expression in bovine skeletal muscle. *BMC genomics* **12**, 23, doi:10.1186/1471-2164-12-23 (2011).
26 Kim, R. S., Ji, H. & Wong, W. H. An improved distance measure between the expression profiles linking co-expression and co-regulation in mouse. *BMC bioinformatics* **7**, 44, doi:10.1186/1471-2105-7-44 (2006).
27 Marco, A., Konikoff, C., Karr, T. L. & Kumar, S. Relationship between gene co-expression and sharing of transcription factor binding sites in Drosophila melanogaster. *Bioinformatics* **25**, 2473-2477, doi:10.1093/bioinformatics/btp462 (2009).
28 Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737-741, doi:10.1038/nature02046 (2003).
29 Das, J., Mohammed, J. & Yu, H. Genome-scale analysis of interaction dynamics reveals organization of biological networks. *Bioinformatics* **28**, 1873-1878, doi:10.1093/bioinformatics/bts283 (2012).
30 Nepusz, T., Yu, H. & Paccanaro, A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods* **9**, 471-472, doi:10.1038/nmeth.1938 (2012).
31 Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with protein-protein interactions. *Genome research* **12**, 37-46, doi:10.1101/gr.205602 (2002).
32 Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104-110, doi:10.1126/science.1158684 (2008).

33 Esteller, M. Non-coding RNAs in human disease. *Nature reviews. Genetics* **12**, 861-874, doi:10.1038/nrg3074 (2011).
34 Croce, C. M. Causes and consequences of microRNA dysregulation in cancer. *Nature reviews. Genetics* **10**, 704-714, doi:10.1038/nrg2634 (2009).
35 Vergoulis, T. *et al.* TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic acids research* **40**, D222-229, doi:10.1093/nar/gkr1161 (2012).
36 Hsu, S. D. *et al.* miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic acids research* **42**, D78-85, doi:10.1093/nar/gkt1266 (2014).
37 Feldman, I., Rzhetsky, A. & Vitkup, D. Network properties of genes harboring inherited disease mutations. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 4323-4328, doi:10.1073/pnas.0701722105 (2008).
38 Goh, K. I. *et al.* The human disease network. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 8685-8690, doi:10.1073/pnas.0701361104 (2007).
39 Lu, M. *et al.* An analysis of human microRNA and disease associations. *PloS one* **3**, e3420, doi:10.1371/journal.pone.0003420 (2008).
40 Jiang, Q. *et al.* miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic acids research* **37**, D98-104, doi:10.1093/nar/gkn714 (2009).
41 Chen, C. Y., Chen, S. T., Fuh, C. S., Juan, H. F. & Huang, H. C. Coregulation of transcription factors and microRNAs in human transcriptional regulatory network. *BMC bioinformatics* **12 Suppl 1**, S41, doi:10.1186/1471-2105-12-S1-S41 (2011).
42 Lin, C. C. *et al.* Crosstalk between transcription factors and microRNAs in human protein interaction network. *BMC systems biology* **6**, 18, doi:10.1186/1752-0509-6-18 (2012).
43 Yu, X., Lin, J., Zack, D. J., Mendell, J. T. & Qian, J. Analysis of regulatory network topology reveals functionally distinct classes of microRNAs. *Nucleic acids research* **36**, 6494-6503, doi:10.1093/nar/gkn712 (2008).
44 Dang, C. V. MYC on the path to cancer. *Cell* **149**, 22-35, doi:10.1016/j.cell.2012.03.003 (2012).
45 Koh, C. M. *et al.* Myc enforces overexpression of EZH2 in early prostatic neoplasia via transcriptional and post-transcriptional mechanisms. *Oncotarget* **2**, 669-683 (2011).
46 Gurel, B. *et al.* Nuclear MYC protein overexpression is an early alteration in human prostate carcinogenesis. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* **21**, 1156-1167, doi:10.1038/modpathol.2008.111 (2008).
47 Chang, T. C. *et al.* Widespread microRNA repression by Myc contributes to tumorigenesis. *Nature genetics* **40**, 43-50, doi:10.1038/ng.2007.30 (2008).
48 Cairns, P. *et al.* Frequent inactivation of PTEN/MMAC1 in primary prostate cancer. *Cancer research* **57**, 4997-5000 (1997).

49 Liang, H. & Li, W. H. MicroRNA regulation of human protein protein interaction network. *RNA* **13**, 1402-1408, doi:10.1261/rna.634607 (2007).
50 Xu, J. *et al.* MiRNA-miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features. *Nucleic acids research* **39**, 825-836, doi:10.1093/nar/gkq832 (2011).
51 Das, J. & Yu, H. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC systems biology* **6**, 92, doi:10.1186/1752-0509-6-92 (2012).
52 Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Human mutation* **21**, 577-581, doi:10.1002/humu.10212 (2003).
53 Stenson, P. D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome medicine* **1**, 13, doi:10.1186/gm13 (2009).
54 Amberger, J., Bocchini, C. & Hamosh, A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM(R)). *Human mutation* **32**, 564-567, doi:10.1002/humu.21466 (2011).
55 Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic acids research* **37**, D793-796, doi:10.1093/nar/gkn665 (2009).
56 Yu, H., Jansen, R., Stolovitzky, G. & Gerstein, M. Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications. *Bioinformatics* **23**, 2163-2173, doi:10.1093/bioinformatics/btm291 (2007).

.

CHAPTER 2

DISSECTING DISEASE INHERITANCE MODES IN A 3D PROTEIN NETWORK CHALLENGES THE "GUILT-BY-ASSOCIATION" PRINCIPLE

Originally published as: Dissecting disease inheritance modes in a 3D protein network challenges the "guilt-by-association" principle. Guo Y, Wei X, Das J, Grimson A, Lipkin SM, Clark AG and Yu H. *Am J Hum Genet* 2013 July 11 3(1): 78-89. Wei X performed the Y2H assay. Das J performed the protein docking. I performed the rest of the analysis and co-wrote the paper with Yu H.

2.1. Abstract

To better understand different molecular mechanisms through which mutations lead to various human diseases, we classified 82,833 disease-associated mutations according to their inheritance modes (recessive vs. dominant) and molecular types [in-frame (missense point mutations and in-frame insertions/deletions) vs. truncating (nonsense mutations and frameshift insertions/deletions)], and systematically examined the effects of different classes of disease mutations in a three-dimensional protein interactome network with the atomic-resolution interface resolved for each interaction. We find that although recessive mutations on the interaction interface of two interacting proteins tend to cause the same disease, this widely-accepted "guilt-by-association" principle does not apply to dominant mutations. Furthermore, recessive truncating mutations on the same interface are much more likely to cause the same disease, even if they are close to the N-terminus of the protein. On the contrary, dominant truncating mutations tend to be enriched between interfaces. These results suggest that a significant fraction of truncating mutations can generate functional protein products. For example, TRIM27, a known cancer-associated protein, interacts with three proteins (MID2, TRIM42, and SIRPA) through three different interfaces. A dominant truncating mutation (p.Tyr342Thrfs*30 [c.1024delT]) associated with ovarian carcinoma is localized between the last two interfaces; the mutated protein retains its interaction with MID2 and TRIM42 through the first two interfaces, but loses its interaction with SIRPA through the third interface. Our findings will help better understand molecular mechanisms of thousands of disease-associated genes and their tens of thousands of mutations, especially for those carrying truncating

mutations, which are often erroneously considered as "knock-out" alleles.

2.2. Introduction

Understanding genotype-to-phenotype relationships has been a central theme of human genetics[1]. In the past few decades, great progress has been made in identifying and characterizing disease-associated genes underlying many Mendelian disorders[2,3]. Advances in next-generation sequencing technologies and genome-wide association studies have further facilitated the identification of allelic variants associated with complex genetic diseases[4,5]. However, it is often unclear how these mutations translate into complex disease phenotypes. Furthermore, disease-associated mutations can be classified into different categories based on their inheritance modes (dominant or recessive) and molecular types (missense or nonsense). Mutations in different categories may cause disease through completely different mechanisms at the molecular level (e.g., loss of function or gain of function)[3,6].

Given that the cell functions as an intricate molecular network, disease mutations not only cause aberrations of single genes, but could also introduce perturbations to the broader network that lead to the observed phenotype[7-10]. Various network-based approaches have been employed to explore genotype-to-phenotype relationships[7-11]. Goh et al. (2007) and Feldman et al. (2008) found that protein products of genes associated with similar diseases are more likely to physically interact, forming disease-specific functional modules[8,9]. Based on this commonly-accepted "guilt-by-association" principle[12], many methods have been developed to predict novel disease-associated genes using the protein interactome network[13-15].

However, none of these methods have considered the potential differences in the molecular mechanisms leading to the corresponding disorders for mutations of different inheritance modes and molecular types. Zhong et al. (2009) found that disease mutations could lead to two types of perturbations at the network level: node removal (loss of all known interactions of a protein) or edgetic perturbation (loss of specific interactions of a protein)[11]. They also found that a higher fraction of mutations associated with autosomal dominant diseases are in-frame, tend to be in structural proteins, and are likely to affect exposed residues[11].

Functional consequences of different classes of disease mutations can be better characterized by considering the three-dimensional (3D) structures of proteins. Recent studies have shown that incorporating structural information with the protein-protein interaction network provides mechanistic understanding of disease-associated genes and mutations at the molecular level[6,16]. In a previous study, we established a high-quality 3D protein interactome network with structurally resolved interfaces for each interaction[16]. We analyzed in-frame disease mutations within this 3D interactome network and found that disease specificity of in-frame mutations can be explained by their locations within corresponding interaction interfaces[16].

However, to date, no systematic analysis has been done to examine the widely-used "guilt-by-association" principle on disease mutations with different inheritance modes and molecular types, which is the focus of this study. We compiled a comprehensive set of disease-associated mutations from the Human Gene Mutation Database (HGMD)[17,18] and the Catalogue of Somatic Mutations in Cancer (COSMIC)[19,20]. We then annotated the inheritance modes of disease mutations based

on manually curated inheritance information. We have further expanded the 3D protein interactome network with 668 additional high-quality binary interactions[16]. Structural details of protein interactions provide a tool to examine the effects of different types of disease mutations at atomic resolution. Here, we applied this approach to systematically analyze disease mutations of different inheritance modes and molecular types, which can be divided into four categories (i.e., dominant in-frame, dominant truncating, recessive in-frame, and recessive truncating). First, we examined the location distribution of disease mutations in different categories on proteins, with respect to interaction interfaces. We then investigated to what extent the "guilt-by-association" principle can be applied to pairs of mutations in different categories at different locations of corresponding proteins. We found that the "guilt-by-association" principle does not apply to dominant disease mutations (both in-frame and truncating). Furthermore, we found that 61% of recessive truncating mutation pairs on the same interaction interface cause the same disease, significantly higher than those on different interfaces (12%). This analysis was not performed in our previous study[16] and our results indicate that a significant fraction of truncating mutations can generate protein products that retain at least some of the wildtype functions, contrary to the common belief that truncating mutations are often complete loss-of-function mutations[21-24].

2.3. Materials and Methods

2.3.1. Compiling a high quality list of disease-associated genes and mutations

Somatic mutations and their associated cancers were obtained from COSMIC[19,20]

(version 56). To remove putative passenger mutations, only mutations on genes in the Cancer Gene Census were included[25-28]. Germline mutations and their associated diseases were obtained from HGMD[17,18] (professional version 2010.12). Only "disease-causing mutations" and "disease-associated polymorphisms of functional significance" were selected for further analyses. Each mutation and its flanking sequence was translated into amino acid and mapped onto the corresponding protein sequence. Protein sequences used were obtained from SwissProt[29] (release 57.6).

The nomenclatures of diseases are not standardized between the two databases. We have compiled a comprehensive disease-gene association map based on the Online Mendelian Inheritance in Man (OMIM)[2] and HGMD databases with unique disease IDs for each phenotypically distinct disorder. To standardize the nomenclature, all disease names were mapped to our disease IDs through bioinformatic processing and manual curation.

2.3.2. Constructing the three-dimensional protein interactome network

The human 3D protein interactome network was constructed as previously described in Wang et al. (2012)[16]. Since the publication, the 3D protein interactome network has been continuously updated[30]. New binary protein interactions have been incorporated[31]. Furthermore, in Wang et al.[16], binary protein interactions that are supported by only one co-crystal structure with no other literature evidence were excluded from the 3D protein interactome network to ensure quality. Here, we modified our filtering criteria to include all binary protein interactions supported by co-crystal structures in 3did[32] or iPfam[33], as co-crystal structures are usually

considered as gold-standard evidence that these interactions exist. Our homology-modeling approach assigns each interface domain with specific interactions of that protein and one interaction could have multiple interface domains. If two proteins interact through multiple domains, all domains involved in the interaction are considered to be the interaction interface. If each of the two interacting proteins has other domains that interact with other proteins, these different domains will be classified as different interfaces for different interactions. To evaluate the performance of our homology-modeling approach, we carried out three-fold cross-validation using the 1,456 human interaction pairs with known co-crystal structures. We found that over 94% of these interactions are correctly predicted with corresponding interaction interfaces, indicating the high accuracy of our approach[16]. Currently, our homology-modeling approach could not account for protein-peptide interactions, as it is extremely difficult to predict protein-peptide interactions with high accuracy[34].

2.3.3. Annotating the inheritance modes

Inheritance information of disease-associated genes was obtained from two sources: Zhong et al. (2009)[11] and the Cancer Gene Census[35]. Each unique gene-disease pair was assigned either autosomal dominant or autosomal recessive inheritance. Gene-disease pairs with other inheritance patterns, e.g. sex-linked inheritance, were discarded. Gene-disease pairs with conflicting annotations in the two datasets were removed. In total, we have collected inheritance patterns for 1,794 unique gene-disease pairs. Next, we separated mutations into either autosomal dominant or autosomal recessive inheritance based on the genes they are on and disease they are

associated with. A total of 38,497 disease-associated mutations with either autosomal dominant or autosomal recessive inheritance were obtained.

2.3.4. Statistical analysis

Mutation distribution with respect to interaction interfaces

Proteins with at least one mutation and at least one interaction domain were chosen for the mutation enrichment calculation. Each protein sequence was divided into three regions, "in interaction interface", "in other domain", and "outside domains". The total number of amino acids and the total number of mutations in each region were counted. If mutations were randomly distributed, the fraction of mutations in reach region should be proportional to the relative length of each region. The expected fraction of mutations in each region (p_2) was calculated by dividing the sum of sequence length of each region in all proteins by the sum of total sequence length of all proteins. The observed fraction of mutations in each region (p_1) was calculated by adding mutations in each region of all proteins and dividing the sum by the total number of mutations. The odds ratios were calculated based on these expected and observed fractions.

$$OR = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)}$$

The 95% confidence interval for the odds ratio[36] and *Z*-scores were calculated by:

$$SE_{logodds} = \sqrt{\frac{1}{n_{mut,region}} + \frac{1}{n_{mut,total} - n_{mut,region}} + \frac{1}{n_{res,region}} + \frac{1}{n_{res,total} - n_{res,region}}}$$

$$95\%\ CI_{\log odds} = \ln OR \pm (N_{0.975} \times SE_{logodds})$$

$$Z = \frac{\ln(OR)}{SE_{logodds}}$$

where $N_{0.975}$ is the 97.5 percentile value of the standard Normal distribution, n_{mut} is the number of mutations, and n_{res} the total number of residues.

Locus heterogeneity calculations

Mutation pairs on interaction partners

Genes with at least one interaction interface and at least one mutation on an interaction interface were selected for this calculation. Of all mutations in these genes, only mutations on interaction interfaces were used, and mutations outside of interaction interfaces were discarded. From this list of genes, all possible pairs of genes where at least one of the two genes has more than one interaction interface were selected.

For each gene pair, all possible mutations pairs where one mutation is on an interaction interface of gene A and another mutation is on an interaction interface of gene B were considered. We have divided all mutation pairs into three categories: if gene A and gene B interact in the 3D protein network and both mutations are in the interaction interface responsible for the interaction between gene A and gene B, the mutation pair was considered to be “in the same interface”; if at least one of the two mutations is not in the interaction interface between gene A and gene B, the mutation pair was considered to be “in other interaction interfaces”; finally, if gene A and gene

B do not interact, the mutation pair was considered to be "non-interacting". We then calculated the percentage of mutation pairs that cause the same disease for each category. The statistical significance of the comparisons between categories was evaluated by the cumulative binomial distribution:

$$P(c \geq c_o) = \sum_{c=c_o}^{N} [\frac{N!}{N!(N-c)!}] p^c (1-p)^{N-c}$$

where N is the total number of mutation pairs, c_o is the number of observed pairs causing the same disease, and p is the fraction of pairs causing the same disease in the control sample. In all calculations, the binomial test was performed twice, where the test and the control groups were swapped in the second test. The least significant P-value was used.

Effect of mutation location on locus heterogeneity

Here, all genes with at least one mutation were selected. All possible pairs of genes that interact in the 3D protein network were used. For each gene pair, gene A was divided into three equal parts, and mutations on each third were paired with all mutations on gene B that were on the corresponding interaction interface. All mutation pairs were classified into two categories: if the mutation on gene A is also on the corresponding interaction interface with gene B, the mutation pair was considered to be on the "interaction interface", otherwise, the mutation pair was classified as "other". For each gene pair, both genes were used once as gene A and once as gene B. The statistical significance of the difference between the two categories for each third was evaluated by the cumulative binomial distribution as described above.

Enrichment of truncating mutations in inter-domain regions

Here, an inter-domain region was defined as a region between two different interaction interfaces on a protein. Two interaction interfaces on a protein were considered different if there exists at least one protein that interacts with one interface but not the other in the 3D protein interaction network. Genes with at least one inter-domain region and at least one truncating mutation were selected for further analyses. The enrichment of dominant and recessive truncating mutations in the inter-domain regions was measured by an odds ratio as detailed above. Sample sizes in all the calculations are listed in Table B.2.

2.3.5. Identification of loss-of-function dominant mutations

Huang et al. (2010)[37] have made a genome-wide prediction of the probability of genes to exhibit haploinsufficiency. Based on their predicted probability scores of being haploinsufficient [*p*(HI)], we considered the top 10% genes with the highest *p*(HI) as haploinsufficient genes. To validate the predicted haploinsufficient (HI) gene set, we checked whether the predicted HI genes tend to be dominantly inherited. Among 583 dominantly inherited genes, 161 were predicted to be HI genes, while only 32 out of 515 recessively inherited genes were predicted to be haploinsufficient. The enrichment for dominantly inherited genes in the HI gene set verifies the prediction accuracy. We classified all dominant mutations on HI genes as 'haploinsufficient mutations' and all dominant mutations that are not on HI genes as 'non-haploinsufficient mutations'.

2.3.6. Selection of proof-of-principle example for experimental validation

To experimentally validate our hypothesis that protein products of alleles with truncating mutations located in between interaction interfaces can retain some of their original functions/interactions, we searched for dominant truncating mutations that satisfy the following criteria: 1) The mutation has to be located between two different interaction interfaces. 2) Dominant truncating mutations have to be enriched in the inter-domain region of the gene. 3) To test the conservation/loss of specific interactions of the truncated protein with yeast two-hybrid (Y2H) assays, the interactions between the wildtype protein and its interactors at different interaction interfaces must be detectable by our Y2H pipeline. TRIM27 (MIM 602165) p.Tyr342Thrfs*30 (c.1024delT) was chosen for experimental validation because it satisfies all the above requirements and we have existing clones of *TRIM27*, *SIRPA* (MIM 602461), *MID2* (MIM 300204), and *TRIM42*.

2.3.7. Determination of interaction interfaces using the 3D protein interaction network and structural interface matching

Using a combination of the 3D protein interaction network and structural interface matching, we determined domains mediating interactions between TRIM27-SIRPA, TRIM27-MID2, and TRIM27-TRIM42. Structural interface matching comprised of two steps – rigid body docking and flexible docking[38]. For putative interacting domains, crystal structures were obtained from PDB[39]. Only high-resolution (< 2.5Å) x-ray diffraction structures were used. Rigid body docking was performed using Patchdock[40,41] with default parameters. This was followed by backbone refinement of

the two proteins using normal mode analysis[42]. Finally, both the side chain and backbone conformations were refined using the computationally efficient FiberDock algorithm with default parameters[43,44]. We find that a SPRY domain on TRIM27 and a C1-set domain on SIRPA mediate the interaction between TRIM27 and SIRPA. An energetically feasible solution was found by docking the SPRY domain (PDB id: 2YYO, crystallized by the RIKEN Structural Genomics/Proteomics Initiative, yet to be published) and the C1-set domain (PDB id: 2WNG)[45]. Both the TRIM27-MID2 and TRIM27-TRIM42 interactions are mediated by zf-B box domains on the corresponding proteins. The energetic feasibility of dimerization of the zf-B box domain is demonstrated by a co-crystal structure (PDB id: 2YVR, crystallized by the RIKEN Structural Genomics/Proteomics Initiative, yet to be published).

2.3.8. Construction of plasmids and disease mutant clones

Wildtype *TRIM27, MID2, TRIM42,* and *SIRPA* entry clones are from the hORFeome 3.1 collection[46]. To generate disease mutant clones, PCR mutagenesis was carried out as previously described[11,16,47]. Briefly, wildtype *TRIM27* in an activation domain (AD) vector was used as the template in PCR reactions to generate N- and C-terminal fragments, each containing the desired mutation in their overlapping region. BP recombination reactions were performed according to the manufacturer's manual (Gateway® BP Clonase® II enzyme mix, catalog number 11789-020) to move mutant clones into the entry vector.

2.3.9. Yeast two-hybrid

Y2H was performed as previously described[48]. Briefly, wildtype and mutant *TRIM27* were transferred into an AD vector. Wildtype *MID2, TRIM42,* and *SIRPA* were transferred into a DNA-binding (DB) vector. AD and DB constructs were transformed into yeast two-hybrid strains *MAT***a** Y8800 and *MAT*α Y8930, respectively. Transformed yeast were spotted onto YPD plates and incubated at 30ºC for ~20 hours before replica-plating onto SC plates that lack Leu and Trp. Yeast cells were allowed to grow at 30ºC for 24 hours before replica-plating onto each of the four selection plates (SC-Leu-Trp-His, SC-Leu-His+CYH, SC-Leu-Trp-Ade, and SC-Leu-Ade+CYH). At 72 hours after replicating, plates were evaluated for protein interactions.

2.4. Results

2.4.1. Mapping disease mutations onto the three-dimensional protein interactome network

Here, we have compiled a comprehensive list of human disease mutations, including 68,789 germline mutations on 2,781 genes associated with 2,244 phenotypically distinct Mendelian diseases from HGMD[17,18], and 14,044 somatic cancer mutations on 366 genes associated with 112 cancers from COSMIC[19,20]. As COSMIC includes results from whole-genome sequencing experiments, it is likely that some mutations identified are passenger mutations that are not causal to the cancer phenotype. To remove putative passenger mutations from the COSMIC dataset, we included only mutations on genes in the Cancer Gene Census[35], a literature-curated list of known cancer-associated genes.

Disease mutations can be dominant or recessive at the cellular level. For dominant mutations, a single mutated allele can lead to pathogenesis, whereas for recessive mutations, both alleles need to be mutated for disease to occur. To examine the potential differences between dominant and recessive mutations, we compiled a list of genes with manually curated inheritance and disease information from published datasets[11,35]. Disease mutations were then classified as autosomal dominant or autosomal recessive according to the inheritance mode of the respective gene and the disease they are associated with. In total, we have annotated the inheritance modes of 38,497 disease-associated mutations.

Using our recently developed "homology modeling" approach[16] and incorporating newly published binary protein-protein interactions, we have generated a high-quality 3D atomic-resolution protein interactome network, comprising of 4,890 structurally resolved interactions involving 3,174 proteins (Figure 2.1A). A total of 11,290 dominant mutations and 8,702 recessive mutations were mapped onto their corresponding proteins in the 3D protein interactome network.

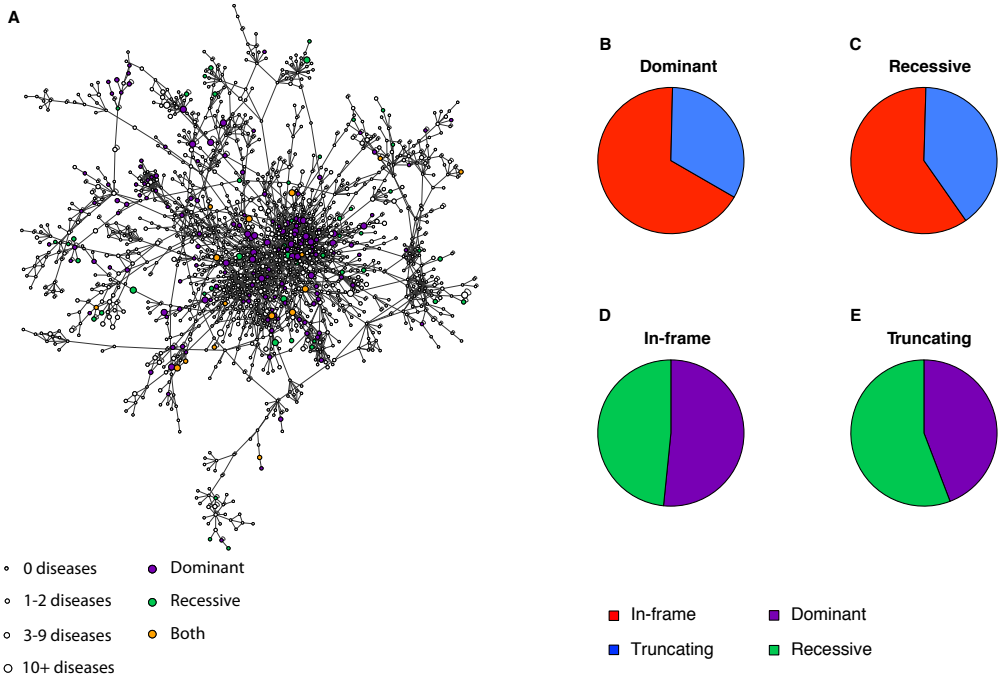


Figure 2.1 Disease-associated genes and mutations in the 3D protein interactome network. **(A)** Network representation of the structurally resolved disease interactome. **(B)** Proportions of in-frame and truncating mutations among all dominant mutations. **(C)** Proportions of in-frame and truncating mutations among all recessive mutations. **(D)** Proportions of dominant and recessive mutations among all in-frame mutations. **(E)** Proportions of dominant and recessive mutations among all truncating mutations.

2.4.2. Different molecular mechanisms between dominant and recessive mutations

All disease mutations can be further divided into two broad classes according to their molecular types and effects on the translated protein products: missense point mutations and in-frame insertions or deletions are classified as in-frame mutations; nonsense mutations and frameshift insertions or deletions are classified as truncating mutations[11,16]. In-frame alleles are likely to produce full-length protein products with local defects, whereas truncating alleles, which are also called "complete loss of function (LoF)" alleles[21], are often assumed not to produce any functional protein

products especially in many current whole-exome and whole-genome sequencing studies[22-24,49].

Among the dominant mutations, 67% are in-frame, while 60% of the recessive mutations are in-frame (Figure 2.1B-C). Conversely, 52% of the in-frame mutations are dominant whereas only 44% of the truncating mutations are dominant (Figure 2.1D-E). The results agree with our current knowledge of the mechanisms of action of dominant mutations. Other than the case of haploinsufficiency, dominance is most often a result of gain-of-function mutations or dominant negative mutations, where the mutated protein product is activated for a specific function or interferes with the normal function(s) of the wildtype protein[11,50]. Therefore, most dominant mutations should translate into specific localized changes on the protein, which can be more easily achieved with in-frame mutations than with truncating mutations.

Most proteins carry out their functions through interactions with other proteins. Our recent study has demonstrated that disruption of specific interactions of a protein is an important mechanism for pathogenesis of many human disease-associated genes and their mutations[16]. To further investigate the differences between dominant and recessive disease mutations, we mapped the mutations onto their corresponding proteins in the 3D protein interactome and examined their locations with respect to interaction interfaces and other functional domains of the proteins. We found that for recessive mutations, both in-frame and truncating mutations are significantly enriched on interaction interfaces (Odds ratio = 3.3, $P < 10^{-20}$ by Z-test; Odds ratio = 1.9, $P < 10^{-20}$ by Z-test; respectively, Figure 2.2B). As recessive mutations are more likely to be loss-of-function mutations[6,51], our results demonstrate that the disruption of protein

interaction interfaces is a common mechanism leading to the loss of specific functions. Dominant in-frame mutations are also enriched on interaction interfaces (Odds ratio = 1.5, $P < 10^{-20}$ by Z-test, 2.2A). As a significant fraction of dominant mutations are gain-of-function mutations[6,51], the results suggest that changes on protein interaction interfaces not only lead to the loss of specific interactions, but also have the potential to generate new ones. Remarkably, the dominant truncating mutations are enriched outside of functional domains (Odds ratio = 3.6, $P < 10^{-20}$ by Z-test, Figure 2.2A), and are depleted on protein interaction interfaces. This shows that the molecular mechanisms of dominant truncating mutations tend to be distinct from their recessive counterparts. To further assess the differences between dominant and recessive mutations, we next investigated the disease specificity of mutations in different categories.

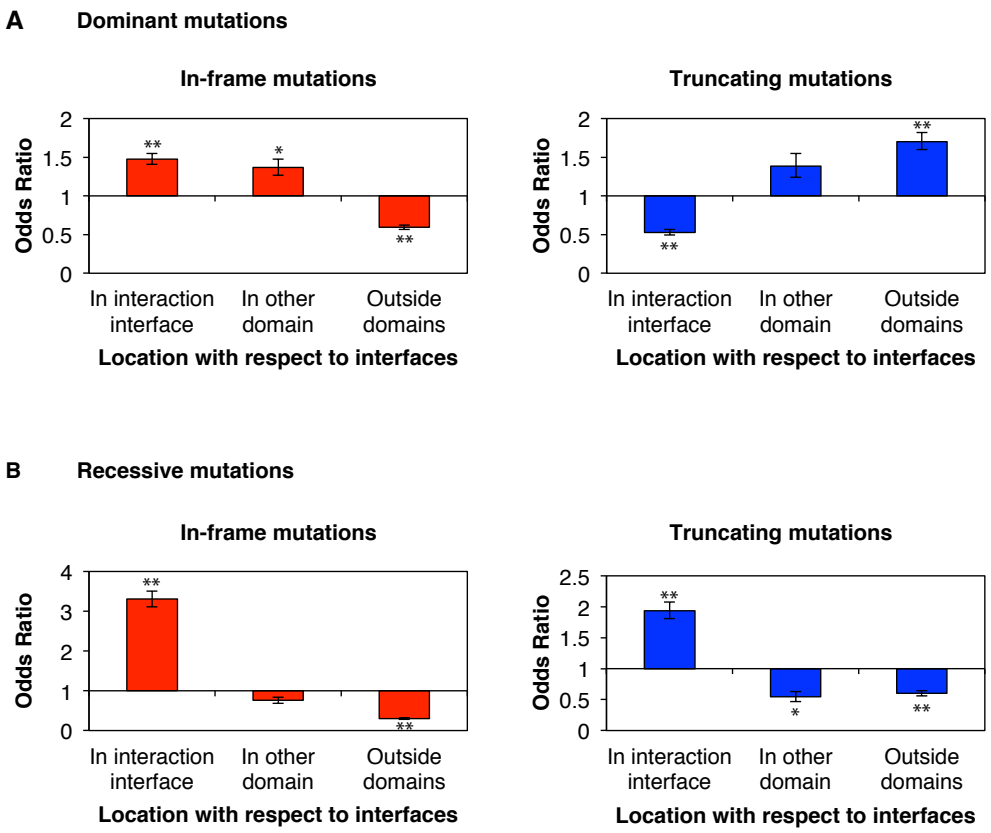


Figure 2.2 Distribution of recessive and dominant disease mutations with respect to interaction interfaces. **(A)** Odds ratios of the distributions of dominant in-frame (left) and truncating (right) mutations on different locations of proteins. **(B)** Odds ratios of the distribution of recessive in-frame (left) and truncating (right) mutations on different locations of proteins. $^{**}P < 10^{-20}$, $^{*}P < 10^{-10}$. *P*-values are calculated using *Z*-tests for log odds ratio. Error bars represent 95% confidence intervals of odds ratios.

2.4.3. "Guilt-by-association" does not apply to dominant mutations

Many genetic diseases show locus heterogeneity, where a disease is associated with mutations on more than one gene. Understanding how different genes converge functionally to associate with the same disorder has important implications in the search for novel disease-associated genes and drug targets. Previous studies have shown that interacting protein pairs are more functionally similar and tend to be associated with the same diseases[12,52]. More specifically, it has recently been shown

that two in-frame mutations on the corresponding interaction interfaces of two interacting proteins tend to cause the same disease[16]. This provides a higher resolution explanation for the "guilt-by-association" principle: mutations on the interaction interface of two interacting proteins disrupt the same interaction in the cellular network, and therefore abolish the same function and cause the same disorder.

To investigate whether "guilt-by-association" holds for both dominant and recessive mutations, we examined the likelihood of in-frame mutation pairs on two different proteins causing the same disease. Among recessive in-frame mutations, 88% of mutation pairs on the corresponding interfaces of two interacting proteins cause the same disease, significantly higher than mutation pairs on interaction interfaces that are not responsible for the interaction between the two proteins [21%, $P < 10^{-20}$ by cumulative binomial test; Figure 2.3A (left)]. In contrast, among dominant in-frame mutations, only 10.1% of mutation pairs on the corresponding interfaces of interacting proteins cause the same disorder. Furthermore, the probability of two dominant in-frame mutations on interacting proteins causing the same disease does not depend on whether the two mutations are located on the corresponding interaction interface responsible for the interaction between the two proteins or on other interaction interfaces [10.1% and 10.6%, respectively; Figure 2.3A (right)]. To further investigate the possible mechanisms of truncating mutations, we repeated the above calculation with truncating mutations. Interestingly, the observed difference between dominant and recessive in-frame mutations can also be seen among truncating mutations (Figure 2.3B). The likelihood of recessive truncating mutation pairs on interacting proteins causing the same disease depends on their location relative to the interaction interface

[Figure 2.3B (left)]. In contrast, dominant truncating mutation pairs on interacting proteins are less likely to cause the same disease, regardless of their location [Figure 2.3B (right)]. Just like in-frame mutations, the "guilt-by-association" principle applies well to recessive truncating mutations, but not to dominant ones.

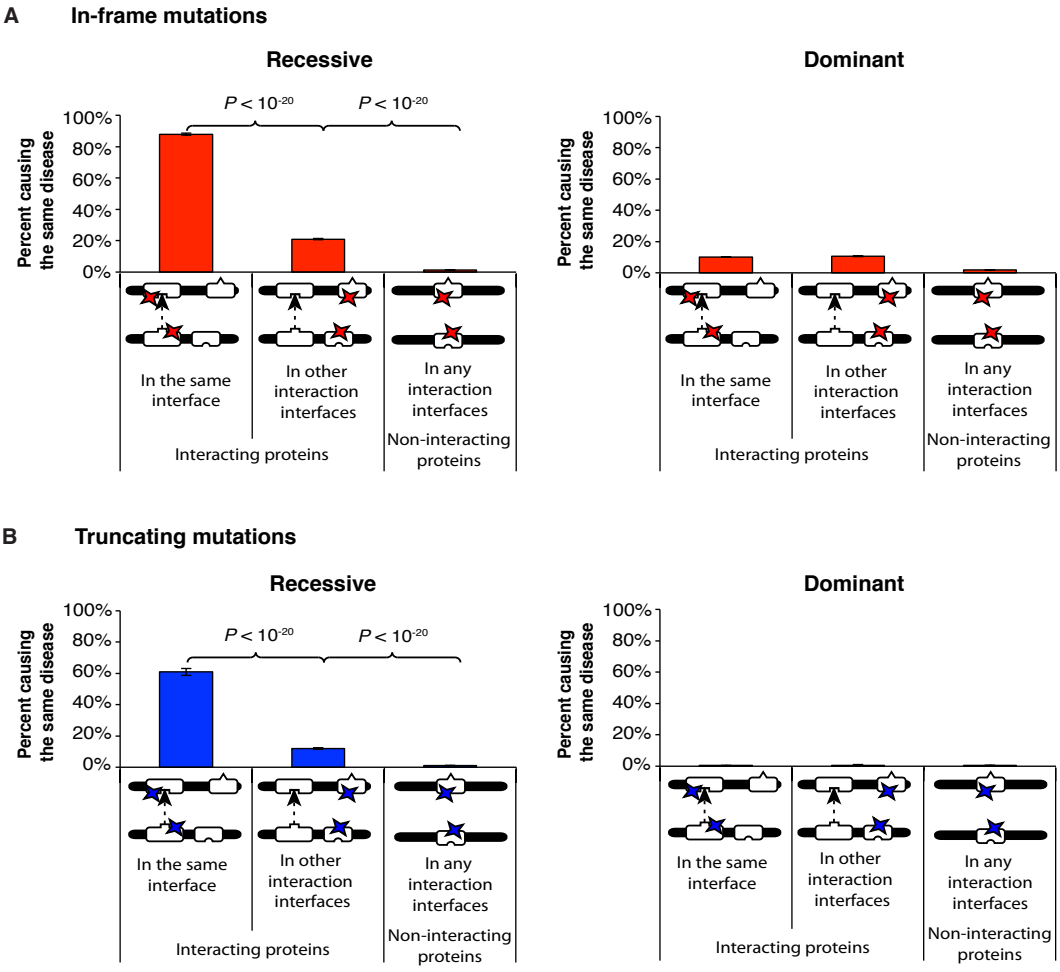


Figure 2.3 Analysis of locus heterogeneity among dominant and recessive disease mutations. **(A)** Percentage of recessive (left) or dominant (right) in-frame mutation pairs on two different proteins causing the same disease. **(B)** Percentage of recessive (left) or dominant (right) truncating mutation pairs on two different proteins causing the same disease. Error bars represent ±SE. *P*-values are calculated using cumulative binomial tests.

An interesting example of "guilt-by-association" can be observed in Glanzmann thrombasthenia (MIM 273800), which is associated with recessive mutations on the corresponding interaction interfaces of both *ITGA2B* (MIM 607759) and *ITGB3* (MIM 173470). On the other hand, dominant mutations on the interaction interface of two proteins are often associated with different diseases. For example, dominant mutations on the Calcium-binding EGF domains of *FBN1* (MIM 134797) are associated with Marfan syndrome (MIM 154700), whereas dominant mutations on the corresponding interaction interface of *FBN2* (MIM 612570) are associated with contractural arachnodactyly (MIM 121050). Although Marfan syndrome and contractural arachnodactyly are related diseases, they have distinct clinical phenotypes[53].

Our "guilt-by-association" analysis demonstrates that while recessive mutations on two different proteins disrupting the same interaction tend to cause the same disorder, the same principle cannot be extended to dominant mutations. A likely explanation for these results is that loss-of-function mutations on two interacting proteins often cause the same disease by disrupting the same edge in the interaction network, but gain-of-function mutations on interacting proteins are less likely to cause the same disease as mutations on two different genes rarely gain the same function. While recessive mutations are more likely to be loss-of-function mutations, dominant mutations can be gain-of-function, dominant negative, or loss-of-function mutations (in the case of haploinsufficiency). To differentiate the molecular mechanisms of different classes of dominant mutations, we have divided all dominant mutations into two categories: those that are likely to cause disease through haploinsufficiency

(haploinsufficient mutations) and those that are not likely to cause disease through haploinsufficiency (non-haploinsufficient mutations) based on a genome-wide prediction of haploinsufficient genes[37]. We found that two haploinsufficient in-frame mutations on the corresponding interaction interfaces between interacting proteins are significantly more likely to cause the same disease compared to two non-haploinsufficient in-frame mutations on the corresponding interfaces of interacting proteins ($P < 10^{-20}$ by cumulative binomial test; Figure 2.4A). Our results support the idea that because a large fraction of dominant mutations are gain-of-function, the "guilt-by-association" principle does not apply to these mutations. A similar calculation could not be performed on truncating mutations due to the small sample size. However, we found a clear distinction in the distribution patterns of haploinsufficient and non-haploinsufficient truncating mutations on their corresponding proteins. Similar to recessive truncating mutations, haploinsufficient truncating mutations are enriched on protein interaction interfaces. In contrast, non-haploinsufficient truncating mutations are highly enriched outside of functional domains (Figure 2.4B). This suggests that truncating mutations can also cause loss or gain of specific functions through distinct molecular mechanisms. Furthermore, the mode of action of truncating mutations can be inferred from their locations with respect to interaction interfaces.

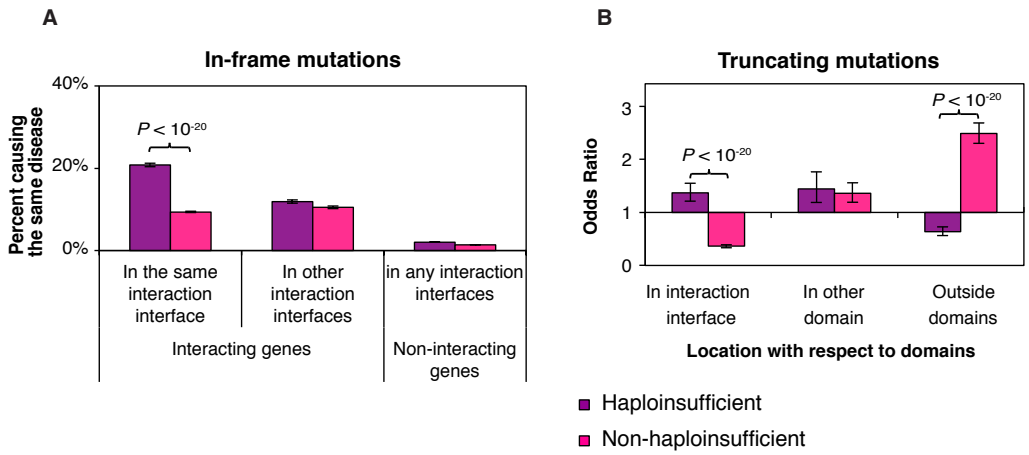


Figure 2.4 (A) Analysis of different molecular mechanisms of dominant mutations. Percentage of haploinsufficient and non-haploinsufficient in-frame mutation pairs on two different proteins causing the same disease. **(B)** Odds ratios of the distribution of haploinsufficient and non-haploinsufficient truncating mutations on different locations of proteins. Error bars represent ±SE. *P*-values are calculated using cumulative binomial tests.

2.4.4. Truncating alleles can give rise to functional products

Currently, truncating mutations are most often regarded as "knock-out" mutations leading to absent or non-functional protein fragments[11,16]. This is because mRNA harboring premature stop codons are known to be selectively degraded by nonsense-mediated mRNA decay (NMD)[54,55], and furthermore, even if the mRNA is translated, the resultant protein fragment is unlikely to fold into a stable product. If most of the truncating mutations lead to loss of protein product, truncating mutations should be randomly distributed across proteins. However, in Figure 2.2 we observed specific enrichment of recessive truncating mutations and specific depletion of dominant truncating mutations on protein interaction interfaces. Figure 2.3 further shows that truncating mutations with different inheritance modes have different patterns of disease association, and that pairs of recessive truncating mutations on the same

interaction interface are much more likely to cause the same disease than those on different interfaces. These results suggest that, contrary to common belief, a significant portion of truncating mutations are translated into functional protein products.

Truncating mutations near the N-terminus delete larger fractions of the wildtype protein. Therefore, it is generally believed that alleles carrying truncating mutations near the N-terminus are even less likely to produce functional products. Here, we investigated the effects of mutation location with respect to the N-terminus on the functional consequences of truncating mutations. We first classified all truncating mutations into three categories: mutations near the N-terminus that truncate more than two thirds of the wildtype protein, mutations near the C-terminus that truncate less than one third of the wildtype protein, and mutations that are located in the middle of the protein that truncate between one third to two thirds of the wildtype protein. Then, for each pair of interacting proteins, we calculated the percentage of truncating mutations in each category that cause the same disease as mutations on the corresponding interaction interfaces of its interaction partner (Figure 2.5 and Figure B.5). If most truncating mutations near the N-terminus cause complete loss of function, all pairs of these mutations should have the same likelihood of causing the same disease, irrespective of whether they are on the same interacting interface or not. However, we found that, regardless of their location relative to the N-terminus, recessive truncating mutations on the corresponding interaction interfaces of two proteins are always more likely to cause the same disease compared to those that are not located on the corresponding interaction interfaces (Figure 2.5). Furthermore, we

also found that irrespective of their location relative to the N-terminus, dominant truncating mutations are always enriched outside of interaction interfaces while recessive truncating mutations are always enriched on the interaction interfaces (Figure B.6). These results show that the location of truncating mutations relative to the N-terminus does not significantly affect the proportion of truncated proteins that retain specific functions. These results, together with our observations in Figures 2.2, 2.3, and 2.4, confirm that a significant fraction of alleles carrying truncating mutations, even those near the N-terminus, can be translated into proteins with specific functions.

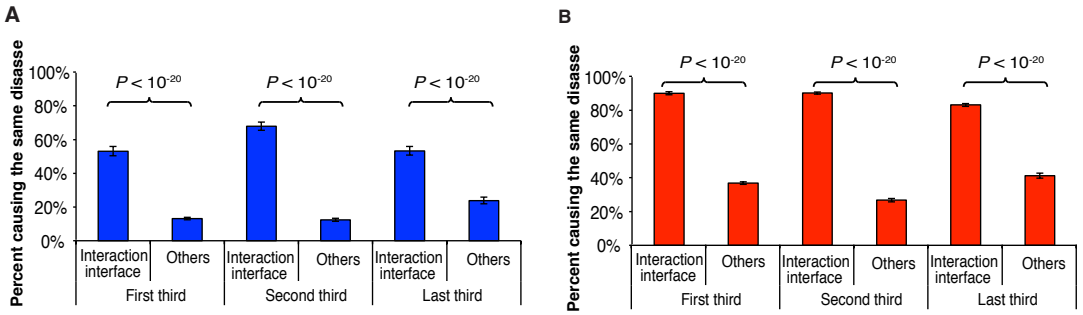


Figure 2.5 Specificity of truncating mutations at different locations of the protein. Percentage of truncating **(A)** or in-frame **(B)** mutations located on different parts of the protein that cause the same disease with mutations on its interaction partner. Error bars represent ±SE. *P*-values are calculated using cumulative binomial tests.

To further characterize the molecular mechanisms underlying dominant and recessive truncating mutations, we calculated the enrichment of disease-associated truncating mutations that occur between two different interaction interfaces. We found that dominant truncating mutations are enriched in sequences located between interaction interfaces (Odds ratio = 1.7, $P < 10^{-20}$ by *Z*-test) while recessive truncating

mutations are depleted in sequences located between interaction interfaces (Odds ratio = 0.74, $P < 10^{-5}$ by *Z*-test; Figure 2.6A; Figure B.7). This result confirms that dominant truncating mutations tend to preserve specific interactions while losing others.

To experimentally validate our conclusions, we tested the interactions of the protein product of *tripartite motif containing 27* (*TRIM27*), a known cancer-associated gene that acts dominantly in oncogenesis[35,56]. A frameshift deletion (p.Tyr342Thrfs*30 [c.1024delT]) occurring just before the SPRY domain of TRIM27 was found to be associated with ovarian carcinoma (MIM 167000)[20,57]. Using a combination of the 3D protein interaction network and structural interface matching, we found that TRIM27 interacts with three other proteins: MID2, TRIM42, and SIRPA, of which only SIRPA interacts exclusively with the SPRY domain on TRIM27 (Figure 2.6B). All three interaction partners of TRIM27 were not previously known to be involved in ovarian cancer. Here, we tested the interactions of wildtype TRIM27 and truncated TRIM27 using Y2H. As the truncating mutation occurs after the interaction interfaces with MID2 and TRIM42 but before the interaction interface with SIRPA, we hypothesized that the truncated TRIM27 would lose its interaction with SIRPA while retaining the other two interactions. The Y2H results confirm that the truncating mutation only disrupts the TRIM27-SIRPA interaction while leaving the other two interactions unaffected (Figure 2.6C). This supports our hypothesis that truncating mutations can retain specific interactions/functions. This result also suggests that abolition of the interaction between TRIM27 and SIRPA might contribute to the cancer phenotype, and SIRPA might be a previously unidentified

ovarian carcinoma-associated gene.

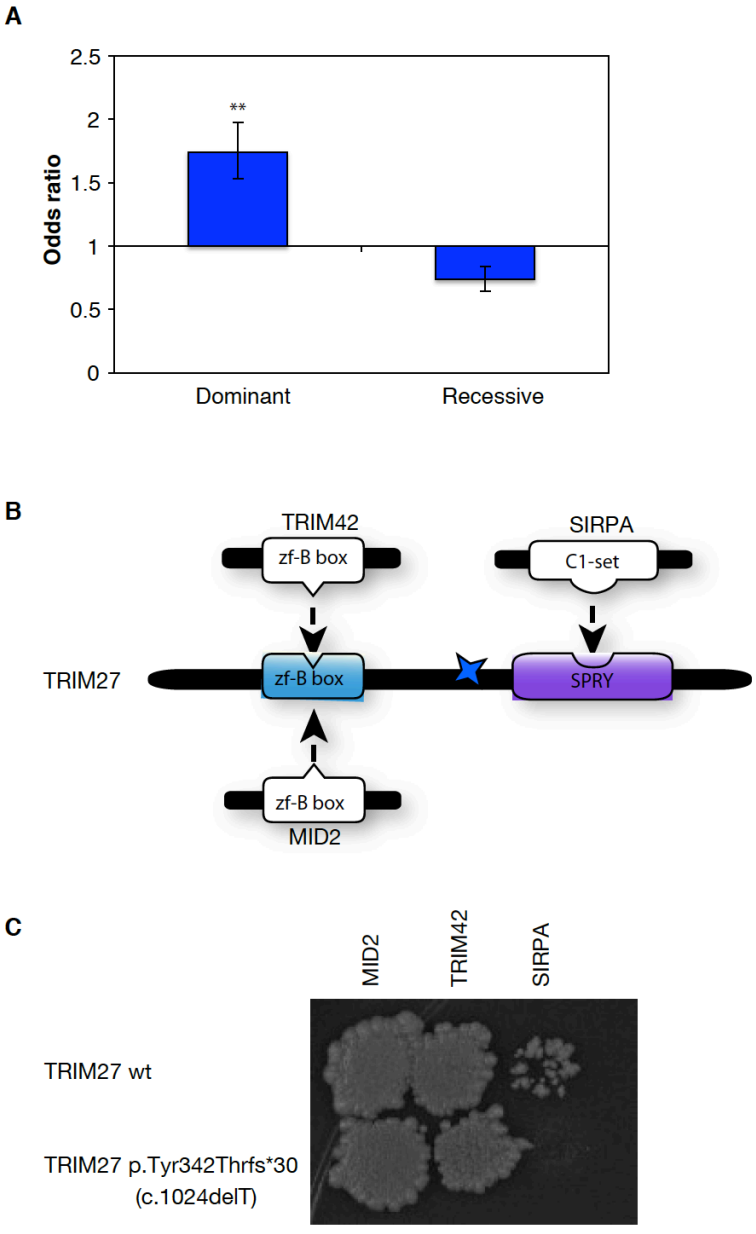


Figure 2.6 Enrichment of truncating mutations between two interaction interfaces. **(A)** Odds ratios of dominant and recessive truncating mutations located between interaction interfaces. $^{**}P < 10^{-20}$, $^{*}P < 0.05$. *P*-values are calculated using *Z*-tests for log odds ratio. Error bars represent 95% confidence intervals of odds ratios. **(B)** Illustration of TRIM27 and its interaction interfaces with MID2, TRIM42, and SIRPA. Colored star indicates the location of the experimentally tested mutation (p.Tyr342Thrfs*30 [c.1024delT]). **(C)** Effects of truncating mutation p.Tyr342Thrfs*30 (c.1024delT) on the interactions of TRIM27 tested by Y2H. Y2H performed by Dr. Xiaomu Wei.

2.5. Discussion

One challenge in deciphering the molecular basis of genetic diseases is that disease phenotypes are often associated with multiple mutations on different genes with variable associated risks. Studies have found that genes associated with the same disease tend to cluster in functional modules within biological networks[8,9]. This "guilt-by-association" principle has been widely applied to identify novel disease-associated genes[58]. However, the accuracy of these predictions is still relatively low[59]. Here, we systematically dissect the "guilt-by-association" principle based on the molecular types and inheritance modes of over 20,000 mutations. While recessive disease mutations on the corresponding interaction interfaces of interacting proteins tend to cause the same disease, the same does not apply to dominant disease mutations. Although current tools that predict disease-associated genes have integrated the protein-protein interactome network with disease phenotypic information to improve the accuracies of predictions[13-15], none of the current prediction models incorporate the difference in inheritance modes of disease-associated genes. By pointing out that the "guilt-by-association" principle only applies for recessive mutations, our findings could significantly improve the accuracy of current prediction methods for disease-associated genes.

Furthermore, truncating mutations, also called LoF mutations, are often regarded as "knock-out" mutations in large-scale mutational screens and genome sequencing projects[21-24,49]. However, there are instances reported where mRNAs harboring truncating mutations escape NMD and are translated into proteins with

dominant negative activities[60,61]. One particularly interesting case study involving *SOX10* (MIM 602229) demonstrated that truncating mutations on different locations of *SOX10* confer distinct neurological phenotypes. Among all *SOX10* alleles harboring nonsense or frameshift mutations, transcripts with mutations on exons 3 and 4 are targeted by NMD, causing a neurological phenotype called Waardenburg-Shah syndrome (MIM 277580). On the other hand, transcripts with mutations in exon 5 escape NMD and lead to a more severe phenotype due to the dominant negative effects of the translated protein[61]. Furthermore, a recent publication revealed that, contrary to common belief, only a small percentage (16.3%) of LoF alleles show significant evidence of NMD[21]. Our results further suggest that it is overly simplistic to consider all truncating mutations as null mutations, as a significant fraction of them do generate functional protein products. Interestingly, our results show that truncating mutations that lead to functional products are not limited to the extreme C-terminal region of the proteins; many proteins can lose more than two thirds of their length and still retain specific functions.

All results that we discussed above are robust to the removal of protein hubs and domain hubs (Figures B.1-B.4), confirming that these results are not biased by over-represented proteins or domain families. Moreover, although filtering COSMIC mutations within Cancer Gene Census genes will enrich for cancer-causing mutations (Figure B.10) and this filtering scheme is often used to select a high-confidence set of cancer mutations[25-28], some of the filtered mutations might still be passenger mutations. Therefore, we repeated our calculations using only the HGMD mutations and all results remain the same (Figures B.8-B.9). These results indicate that, although

cancer is a complex disease, cancer mutations are likely to disrupt normal protein functions through similar biophysical and/or biochemical mechanisms at the molecular level as Mendelian mutations.

In recent years, large numbers of mutations/variants have been discovered from whole genome/exome sequencing studies. Popular tools such as Polyphen-2[62], SIFT[63], and MutationTaster[64] estimate the impact of amino acid substitutions on the respective protein, and are frequently used to prioritize variants discovered from exome sequencing projects. Our method could potentially be used in conjunction with these tools to generate hypotheses regarding the molecular mechanisms of the deleterious variants discovered. Moreover, it might be interesting to consider the penetrance and expressivity of the disease mutations in future analyses[65,66], when sufficient information is available.

In conclusion, by integrating inheritance information with atomic-resolution structural details of protein interactions, our analysis provides an approach to predicting functional consequences at the molecular level for both in-frame and truncating mutations/variants, especially those discovered by various ongoing genome sequencing efforts.

2.6. Acknowledgements

The authors wish to thank Nicolas A Cordero for critical reading of the manuscript. This work was supported by the National Institutes of Health grants R01 GM104424 (to H.Y. and S.M.L.), R01 CA167824 (to H.Y. and S.M.L.), and R01 HG003229 (to A.G.C.); by Weill Cornell Medical College Clinical and Translational Science Center

(CTSC) Pilot Award (to H.Y. and S.M.L.); by Cornell University Seed Grant for Collaborations Between Cornell University-Ithaca and Weill Cornell Medical College Faculty (to H.Y. and S.M.L.); by a donation from Matthew Bell (to S.M.L.); by the Cornell Presidential Life Sciences Fellowship (to Y.G.); and by the Tata Graduate Fellowship (to J.D.). The authors declare that they have no conflict of interest.

REFERENCES

1 Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature genetics* **33 Suppl**, 228-237, doi:10.1038/ng1090 (2003).

2 Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic acids research* **37**, D793-796, doi:10.1093/nar/gkn665 (2009).

3 Jimenez-Sanchez, G., Childs, B. & Valle, D. Human disease genes. *Nature* **409**, 853-855, doi:10.1038/35057050 (2001).

4 Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881-888, doi:10.1126/science.1156409 (2008).

5 McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics* **9**, 356-369, doi:10.1038/nrg2344 (2008).

6 Schuster-Bockler, B. & Bateman, A. Protein interactions in human genetic diseases. *Genome biology* **9**, R9, doi:10.1186/gb-2008-9-1-r9 (2008).

7 Barabasi, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature reviews. Genetics* **12**, 56-68, doi:10.1038/nrg2918 (2011).

8 Feldman, I., Rzhetsky, A. & Vitkup, D. Network properties of genes harboring inherited disease mutations. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 4323-4328, doi:10.1073/pnas.0701722105 (2008).

9 Goh, K. I. *et al.* The human disease network. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 8685-8690, doi:10.1073/pnas.0701361104 (2007).

10 Vidal, M., Cusick, M. E. & Barabasi, A. L. Interactome networks and human disease. *Cell* **144**, 986-998, doi:10.1016/j.cell.2011.02.016 (2011).

11 Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Molecular systems biology* **5**, 321, doi:10.1038/msb.2009.80 (2009).

12 Oliver, S. Guilt-by-association goes global. *Nature* **403**, 601-603, doi:10.1038/35001165 (2000).

13 Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology* **25**, 309-316, doi:10.1038/nbt1295 (2007).

14 Wu, X., Jiang, R., Zhang, M. Q. & Li, S. Network-based global inference of human disease genes. *Molecular systems biology* **4**, 189, doi:10.1038/msb.2008.27 (2008).

15 Wu, X., Liu, Q. & Jiang, R. Align human interactome with phenome to identify causative genes and networks underlying disease families. *Bioinformatics* **25**, 98-104, doi:10.1093/bioinformatics/btn593 (2009).

16 Wang, X. *et al.* Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature biotechnology* **30**, 159-164,

doi:10.1038/nbt.2106 (2012).
17 Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Human mutation* **21**, 577-581, doi:10.1002/humu.10212 (2003).
18 Stenson, P. D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome medicine* **1**, 13, doi:10.1186/gm13 (2009).
19 Forbes, S. A. *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]* **Chapter 10**, Unit 10 11, doi:10.1002/0471142905.hg1011s57 (2008).
20 Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic acids research* **39**, D945-950, doi:10.1093/nar/gkq929 (2011).
21 MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828, doi:10.1126/science.1215040 (2012).
22 Agrawal, N. *et al.* Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science* **333**, 1154-1157, doi:10.1126/science.1206923 (2011).
23 Ernst, T. *et al.* Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. *Nature genetics* **42**, 722-726, doi:10.1038/ng.621 (2010).
24 Seal, S. *et al.* Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nature genetics* **38**, 1239-1241, doi:10.1038/ng1902 (2006).
25 Kaminker, J. S., Zhang, Y., Watanabe, C. & Zhang, Z. CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic acids research* **35**, W595-598, doi:10.1093/nar/gkm405 (2007).
26 Kaminker, J. S. *et al.* Distinguishing cancer-associated missense mutations from common polymorphisms. *Cancer research* **67**, 465-473, doi:10.1158/0008-5472.CAN-06-1736 (2007).
27 Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400-404, doi:10.1038/nature11017 (2012).
28 Pajkos, M., Meszaros, B., Simon, I. & Dosztanyi, Z. Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. *Molecular bioSystems* **8**, 296-307, doi:10.1039/c1mb05246b (2012).
29 The Uniprot Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic acids research* **40**, D71-75, doi:10.1093/nar/gkr981 (2012).
30 Meyer, M. J., Das, J., Wang, X. & Yu, H. INstruct: a database of high quality three-dimensional structurally resolved protein interactome networks. *Bioinformatics*, doi:10.1093/bioinformatics/btt181 (2013).
31 Das, J. & Yu, H. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC systems biology* **6**, 92, doi:10.1186/1752-0509-6-92 (2012).
32 Stein, A., Panjkovich, A. & Aloy, P. 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic acids research*

37, D300-304, doi:10.1093/nar/gkn690 (2009).
33 Finn, R. D., Marshall, M. & Bateman, A. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* **21**, 410-412, doi:10.1093/bioinformatics/bti011 (2005).
34 Petsalaki, E., Stark, A., Garcia-Urdiales, E. & Russell, R. B. Accurate prediction of peptide binding sites on protein surfaces. *PLoS computational biology* **5**, e1000335, doi:10.1371/journal.pcbi.1000335 (2009).
35 Futreal, P. A. *et al.* A census of human cancer genes. *Nature reviews. Cancer* **4**, 177-183, doi:10.1038/nrc1299 (2004).
36 Morris, J. A. & Gardner, M. J. Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *Br Med J (Clin Res Ed)* **296**, 1313-1316 (1988).
37 Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS genetics* **6**, e1001154, doi:10.1371/journal.pgen.1001154 (2010).
38 Tuncbag, N., Gursoy, A., Nussinov, R. & Keskin, O. Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nature protocols* **6**, 1341-1354, doi:10.1038/nprot.2011.367 (2011).
39 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic acids research* **28**, 235-242 (2000).
40 Duhovny, D., Nussinov, R. & Wolfson, H. J. Efficient unbound docking of rigid molecules. *Lect Notes Comput Sc* **2452**, 185-200 (2002).
41 Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. & Wolfson, H. J. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic acids research* **33**, W363-367, doi:10.1093/nar/gki481 (2005).
42 Tirion, M. M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys Rev Lett* **77**, 1905-1908, doi:Doi 10.1103/Physrevlett.77.1905 (1996).
43 Mashiach, E., Nussinov, R. & Wolfson, H. J. FiberDock: a web server for flexible induced-fit backbone refinement in molecular docking. *Nucleic acids research* **38**, W457-W461, doi:Doi 10.1093/Nar/Gkq373 (2010).
44 Mashiach, E., Nussinov, R. & Wolfson, H. J. FiberDock: Flexible induced-fit backbone refinement in molecular docking. *Proteins* **78**, 1503-1519, doi:10.1002/prot.22668 (2010).
45 Hatherley, D., Graham, S. C., Harlos, K., Stuart, D. I. & Barclay, A. N. Structure of Signal-regulatory Protein alpha A LINK TO ANTIGEN RECEPTOR EVOLUTION. *J Biol Chem* **284**, 26613-26619, doi:Doi 10.1074/Jbc.M109.017566 (2009).
46 Lamesch, P. *et al.* hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. *Genomics* **89**, 307-315, doi:10.1016/j.ygeno.2006.11.012 (2007).
47 Suzuki, Y. *et al.* A novel high-throughput (HTP) cloning strategy for site-directed designed chimeragenesis and mutation using the Gateway cloning system. *Nucleic acids research* **33**, e109, doi:10.1093/nar/gni103 (2005).

48 Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104-110, doi:10.1126/science.1158684 (2008).
49 van Haaften, G. *et al.* Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. *Nature genetics* **41**, 521-523, doi:10.1038/ng.349 (2009).
50 Veitia, R. A. Exploring the molecular etiology of dominant-negative mutations. *The Plant cell* **19**, 3843-3851, doi:10.1105/tpc.107.055053 (2007).
51 Lodish, H. F. *Molecular cell biology*. 7th edn, (W.H. Freeman and Co., 2013).
52 Yu, H., Jansen, R., Stolovitzky, G. & Gerstein, M. Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications. *Bioinformatics* **23**, 2163-2173, doi:10.1093/bioinformatics/btm291 (2007).
53 Wang, M., Clericuzio, C. L. & Godfrey, M. Familial occurrence of typical and severe lethal congenital contractural arachnodactyly caused by missplicing of exon 34 of fibrillin-2. *American journal of human genetics* **59**, 1027-1034 (1996).
54 Chang, Y. F., Imam, J. S. & Wilkinson, M. F. The nonsense-mediated decay RNA surveillance pathway. *Annual review of biochemistry* **76**, 51-74, doi:10.1146/annurev.biochem.76.050106.093909 (2007).
55 Maquat, L. E. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nature reviews. Molecular cell biology* **5**, 89-99, doi:10.1038/nrm1310 (2004).
56 Hatakeyama, S. TRIM proteins and cancer. *Nature reviews. Cancer* **11**, 792-804, doi:10.1038/nrc3139 (2011).
57 The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615, doi:10.1038/nature10166 (2011).
58 Wang, X., Gulbahce, N. & Yu, H. Network-based methods for human disease gene prediction. *Briefings in functional genomics* **10**, 280-293, doi:10.1093/bfgp/elr024 (2011).
59 Oti, M., Snel, B., Huynen, M. A. & Brunner, H. G. Predicting disease genes using protein-protein interactions. *Journal of medical genetics* **43**, 691-698, doi:10.1136/jmg.2006.041376 (2006).
60 Fan, S. *et al.* Mutant BRCA1 genes antagonize phenotype of wild-type BRCA1. *Oncogene* **20**, 8215-8235, doi:10.1038/sj.onc.1205033 (2001).
61 Inoue, K. *et al.* Molecular mechanism for distinct neurological phenotypes conveyed by allelic truncating mutations. *Nature genetics* **36**, 361-369, doi:10.1038/ng1322 (2004).
62 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-249, doi:10.1038/nmeth0410-248 (2010).
63 Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* **4**, 1073-1081, doi:10.1038/nprot.2009.86 (2009).
64 Schwarz, J. M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature methods* **7**, 575-576, doi:10.1038/nmeth0810-575 (2010).

65 Niemann, S. & Muller, U. Mutations in SDHC cause autosomal dominant paraganglioma, type 3. *Nature genetics* **26**, 268-270, doi:10.1038/81551 (2000).

66 Zlotogora, J. Penetrance and expressivity in the molecular age. *Genetics in medicine : official journal of the American College of Medical Genetics* **5**, 347-352, doi:10.109701.GIM.0000086478.87623.69 (2003).

CHAPTER 3

GENOME-WIDE ANALYSIS OF EPISTATIC PARTNERS OF HUMAN DISEASE MUTATIONS

Michael Meyer compiled and maintained the protein structures and models, wrote the scripts to identify interface residues, contact residues and surface residues, and wrote the script to identify neighboring residues. Robert Fragoza and Dr. Jin Liang performed the Y2H. Jishnu Das wrote the cherry picking files for the Y2H assays and processed the next generation-sequencing data to identify successfully created mutant clones.

3.1. Introduction

In the past few decades, we have made tremendous progress in our understanding of the genetic bases of many human diseases, in the hope to gain mechanistic understanding of diseases that would allow us to eventually conquer them. Long lists of disease-associated genes and their mutations have been identified to date. However, disease mutations often have varying effects on different individuals, and not all individuals with a specific disease mutation will develop the disease phenotype, i.e. the disease mutation show partial penetrance. The differences in mutation outcome in different individuals can be attributed to several factors, including environmental influence, epigenetic variation and differences in the genetic backgrounds of different individuals. The dependence of mutation outcome on the genetic background is a result of epistatic interactions between different genetic loci. Currently, the molecular mechanisms of epistasis are still not well understood, making epistatic interactions difficult to predict.

The differences in the genetic background between difference species also lead to different mutation outcome for the same mutation on orthologous proteins. For many disease-associated single amino acid substitutions in humans, the deleterious residue in human is the wildtype residue in the orthologous protein of another species. This phenomenon was first observed by Kondrashov et al in 2002, where they proposed that human disease residues could appear as the wildtype in another species due to the presence of coevolved compensatory mutations that neutralize the deleterious effects of these disease residues in the other species[1]. Since this class of disease mutations is hypothesized to be compensated in another species through

epistatic interactions with other mutations, we refer to them as potentially epistastic mutations (PEMs). A recent study found that protein destabilization by a PEM is generally reduced if its neighboring residues in the human protein are substituted to the corresponding residues in a species where the PEM is the wildtype, providing further support to the compensation hypothesis[2].

Computational analyses on the physiochemical properties of PEMs showed that compared to regular disease mutations, PEMs result in less drastic changes in both amino acid volume and hydrophobicity[3]. In addition, PEMs have smaller destabilizing effects on proteins on average compared to regular disease mutations[2]. Interestingly, although PEMs seem to be less deleterious compared to regular disease mutations, they are not more likely to be associated with milder or later onset diseases[4]. Currently, the molecular mechanisms through which PEMs lead to disease are still unclear. To experimentally characterize the molecular phenotypes of PEMs, we systematically examined the impact of PEMs on protein stability and protein-protein interactions using high throughput yeast two-hybrid (Y2H) and GFP assays.

Although examples of compensation have been reported, there has not been a large-scale, systematic study to identify mutations in epistasis with PEMs. Furthermore, previous studies focused mainly on intra-molecular compensation, where the compensatory mutation(s) is on the same protein as the PEM[5,6]. However, it has been shown that PEMs are enriched on the protein surface as well as the protein-protein interaction interface[3,7], suggesting the presence of inter-molecular compensation. In this study, we aim to systematically identify potential epistatic partners of PEMs both intra- and inter-molecularly by searching for amino acid

residues on the same protein or on an interacting protein that coevolve with the PEMs. Identification of residues that are in epistasis with disease mutations can further our understanding on disease penetrance and the molecular mechanisms of the pathological processes of these diseases. Furthermore, epistatic sites identified could be potential drug targets sites for the development of targeted disease therapies.

3.2. Results

3.2.1. Identification and characterization of PEMs

We started with 55,093 missense disease mutations from the Human Gene Mutation Database (HGMD)[8,9]. To identify PEMs, we used the multiple sequence alignment of 99 vertebrate genomes with human generated by the UCSC[10,11] (Figure 3.1). We were able to map 42,005 disease mutations to proteins with an ortholog in at least one other species. Among them, 5,008 disease mutations on 1,058 proteins appear as the wildtype residue in the orthologous protein(s) of least one other species. We consider this 5,008 disease mutations to be PEMs and the rest to be regular, uncompensated disease mutations.

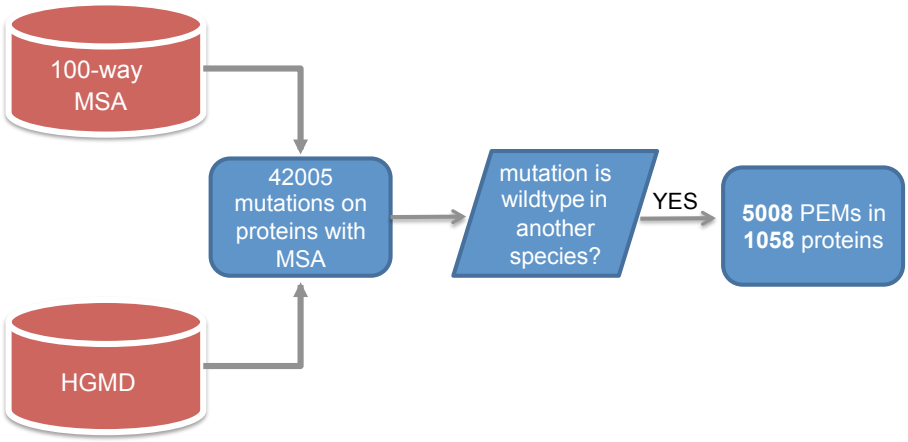


Figure 3.1 Identification of PEMs from HGMD disease mutations and the UCSC 100-way multiple sequence alignment.

As previous physiochemical analyses on PEMs suggest that they are less deleterious on the molecular level[2,3], we would expect the PEMs to be able to exist in the human populations at higher frequencies compared to regular disease mutations. Indeed, using variant data from the Exome Sequencing Project (ESP)[12], we found that 18.8% of PEMs are found as population variants in ESP, while only 5.4% of regular disease mutations are population variants ($P < 10^{-20}$ by cumulative binomial test; Figure 3.2A). Furthermore, we found that compared to regular disease mutations, PEMs tend to have higher allele frequencies (Figure 3.2B). Next, we compared the Polyphen2[13] predictions of the impacts of PEMs and regular disease mutations that are polymorphic in humans, as well as common SNPs with greater than 1% overall allele frequency in ESP. As expected, we found that over 90% of polymorphic regular disease mutations are predicted to be deleterious. On the other hand, only about 45% of polymorphic PEMs are predicted to be deleterious, while 34% of common SNPs are predicted to be deleterious (Figure C.1). Our analyses support the hypothesis that PEMs more benign than regular disease mutations, and further suggest that PEMs are more like common population polymorphisms compared to regular disease mutations.

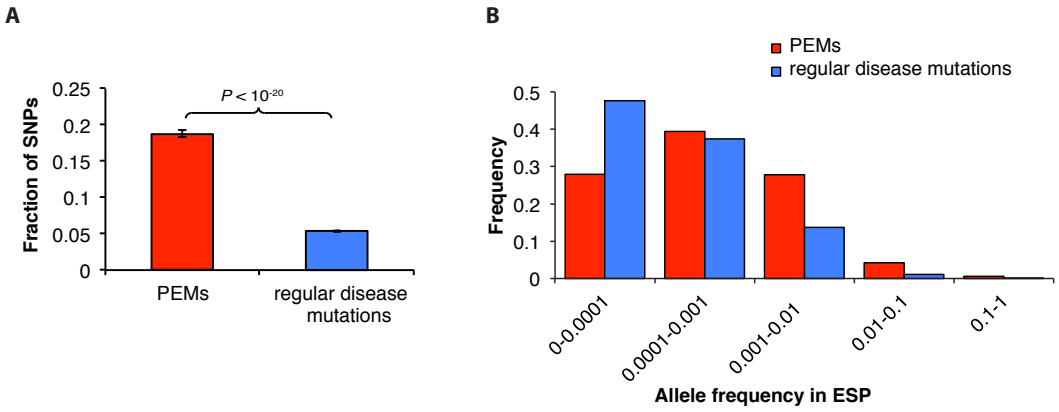


Figure 3.2 Prevalence of PEMs and uncompensated disease mutations in human populations. **(A)** Percentage PEMs and regular disease mutations that exist as population variants. **(B)** Allele frequency distribution of PEMs and regular disease mutations that are polymorphic in humans based on the ESP data.

3.2.2. PEMs could be compensated by inter-molecular epistasis

Previous studies believe that most PEMs are compensated by intra-molecular compensatory mutations[5,6]. However, we have previously found that human disease mutations are enriched on protein-protein interaction interfaces, suggesting that disruption of specific protein-protein interactions is a major mechanism of pathogenesis[14]. Furthermore, it was found that PEMs are enriched on protein surface residues, and also on the protein interaction interface, just like regular disease mutations. It was suggested that the surface/interface residues have fewer intra-protein interactions, making them easier to be compensated for[3,7]. We believe an equally likely hypothesis is that these PEMs on the protein interaction interface are compensated by inter-molecular epistatic mutations on their protein interaction partner(s). Using the 3D protein-protein interaction network generated in our lab based on cocrystal structures, we found that PEMs are not only enriched in the protein interaction interfaces (Odds ratio = 1.97, $P < 10^{-10}$ by Z-test; Figure 3.3A), they are

also equally enriched on interacting domains (Odds ratio = 2.1, $P < 10^{-20}$ by *Z*-test; Figure 3.3A) while depleted in other parts of the protein (Odds ratio = 0.43, $P < 10^{-20}$ by *Z*-test; Figure 3.3A). The level of enrichment of PEMs on protein interaction interfaces and protein interacting domains are similar to that observed in regular disease mutations[15]. This observation suggests that PEMs can cause disease through disruption of protein-protein interactions by changing the conformation of the protein interacting interface. Therefore, it is likely that the interaction disruption can be compensated for by inter-protein epistatic mutations.

Next, we explored the inter-molecular coevolutionary relationships between amino acid residue pairs on interacting proteins. For 4,069 structurally resolved protein interactions with cocrystal structures, we calculated the inter-molecular coevolution scores of all amino acid residue pairs on interacting proteins using Direct Couple Analysis (DCA)[16], by concatenating the multiple sequence alignments of the two interacting proteins. To ensure the accuracy of the predictions, only proteins with at least 50 orthologs are used in this analysis. We took the highest DCA score of each residue as its inter-molecular coevolution score, and we consider residues with the top 5% highest inter-molecular coevolution score as inter-molecularly coevolved residues. Consistent with our expectations, amino acid residues at the protein interaction interface, residues in contact with the interacting protein, residues on the interacting domains, and residues on the protein surface are all enriched with inter-molecularly coevolved residues (Odds ratio = 1.49, 1.39, 1.36 and 1.38 respectively, $P < 10^{-20}$ for all four categories by *Z*-test; Figure 3.3B; see Materials and Methods). In contrast, inter-molecularly coevolved residues are not enriched in regular disease mutations and

SNPs sites (Odds ratio = 1.02 and 1.14 respectively; P= 0.78 and 0.002 respectively by Z-test; Figure 3.3B). Most interestingly, PEMs have the highest enrichment of inter-molecular coevolution among all categories examined here (Odds ratio = 3.26; $P < 10^{-20}$ by Z-test; Figure 3.3B). The results show that PEMs tend to coevolve with amino acid residues on their protein interaction partners, and strongly suggest that many PEMs are compensated by inter-molecular compensatory mutations. The results remain unchanged at different DCA score cutoffs (Figure C.2), and the enrichment of coevolved residues in PEM sites is not due to higher conservation of PEMs compared to other residue categories (Figure C.3).

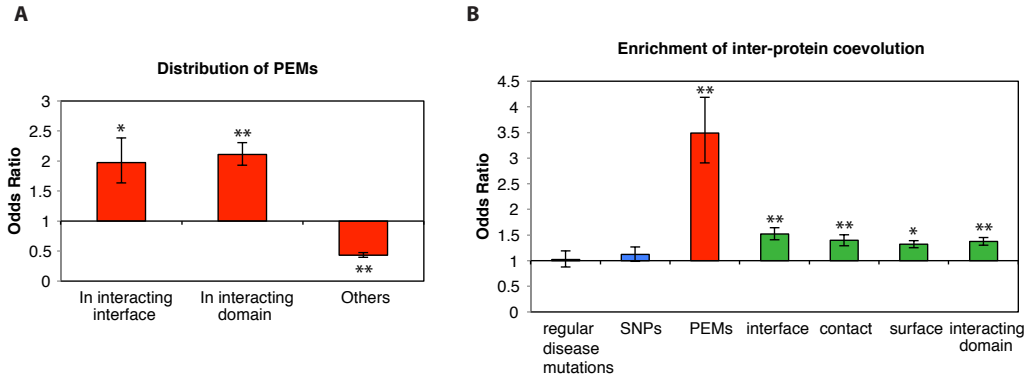


Figure 3.3 Evidence of inter-molecular compensation of PEMs. **(A)** Odds ratios of the distribution of PEMs on different locations of proteins. **(B)** Odds ratios of the enrichment in inter-molecular coevolution at different categories of amino acid sites. $^{**}P < 10^{-20}$, $^{*}P < 0.05$.

3.2.3. Molecular phenotypes of PEMs

In this study, we aim to experimentally characterize the molecular phenotypes of PEMs. We hypothesize that PEMs with inter-molecular compensation are more likely to cause disease through disrupting protein-protein interactions, while PEMs with

intra-molecular compensation are more likely to cause disease by affecting the stability of the protein. Therefore, we would like to test the impacts of candidate inter-molecular PEMs on protein-protein interactions, and the impacts of candidate intra-molecular PEMs on protein stability.

We used both global and local sequence conservation cutoffs to select a list of high-quality candidate inter- and intra-molecular PEMs to test experimentally (Figure 3.4A). One key assumption of the compensation hypothesis is that the human protein with PEM and its orthologous protein in the other species must be functionally equivalent in the two species. Another assumption for the inter-molecular compensation hypothesis is that the interaction between the PEM protein and the compensatory protein is conserved between the two species, which is often not the case between very divergent species. To ensure a reasonable degree of protein function and interaction conservation, we only considered PEMs of which the corresponding orthologs have global sequence identities greater than 70%. As intra-molecular compensation is mostly likely to occur between residues that are in close contact with each other, we examined the neighboring residues of the PEMs in 3D. Protein crystal structures were obtained from the Protein Data Bank (PDB)[17]. For proteins without crystal structures, we then attempt to find a high quality 3D model for it in ModBase[18]. Only high-quality ModBase models with ModPipe Quality Score (MPQS) above 1.1 were used. Next, we considered all residues with Calpha within 8Å to the Calpha of a PEM as neighboring residues of the PEM. This is because 7Å alpha-carbon distance approximately corresponds to a weak hydrogen bond distance between two amino acids[2], and here we used a slightly more lenient cutoff of 8Å.

Then, we consider PEMs with at least one change in its neighboring residues as intra-molecular PEMs, and all neighboring residues that differ between human and the other species with the PEM are consider candidate intra-molecular compensatory mutations. PEMs with fewer or no neighboring residue changes are more likely to be compensated by inter-molecular compensatory mutations. Therefore, we consider PEMs that have greater than 90% local conservation in their neighboring residues as candidate inter-molecular PEMs. In total, we found 394 inter-molecular PEMs on 163 proteins, and 979 intra-molecular PEMs on 318 proteins.

We found that a significantly higher fraction of the inter-molecular PEMs are on the protein surface compared to the intra-molecular PEMs (82% vs. 77%, $P < 10^{-5}$ by Fisher's Exact test, Figure 3.4B). For PEMs on proteins with available cocrystal structures, DCA scores were calculated for all intra-protein amino acid pairs, as well as all inter-protein amino acid pairs between the two monomers of the cocrystal structure. For each residue, we took its highest intra-molecular DCA score as its intra-molecular coevolution score, and its highest inter-molecular DCA score as its inter-molecular coevolution score. We found that a higher fraction of our candidate inter-molecular PEMs have higher inter-molecular coevolution score compared to the candidate intra-molecular PEMs (41% vs. 29%; $P = 0.028$ by Fisher's Exact test, Figure 3.4C). These results show that the inter- and intra-molecular PEMs identified by our method are enriched for the correct type of PEMs.

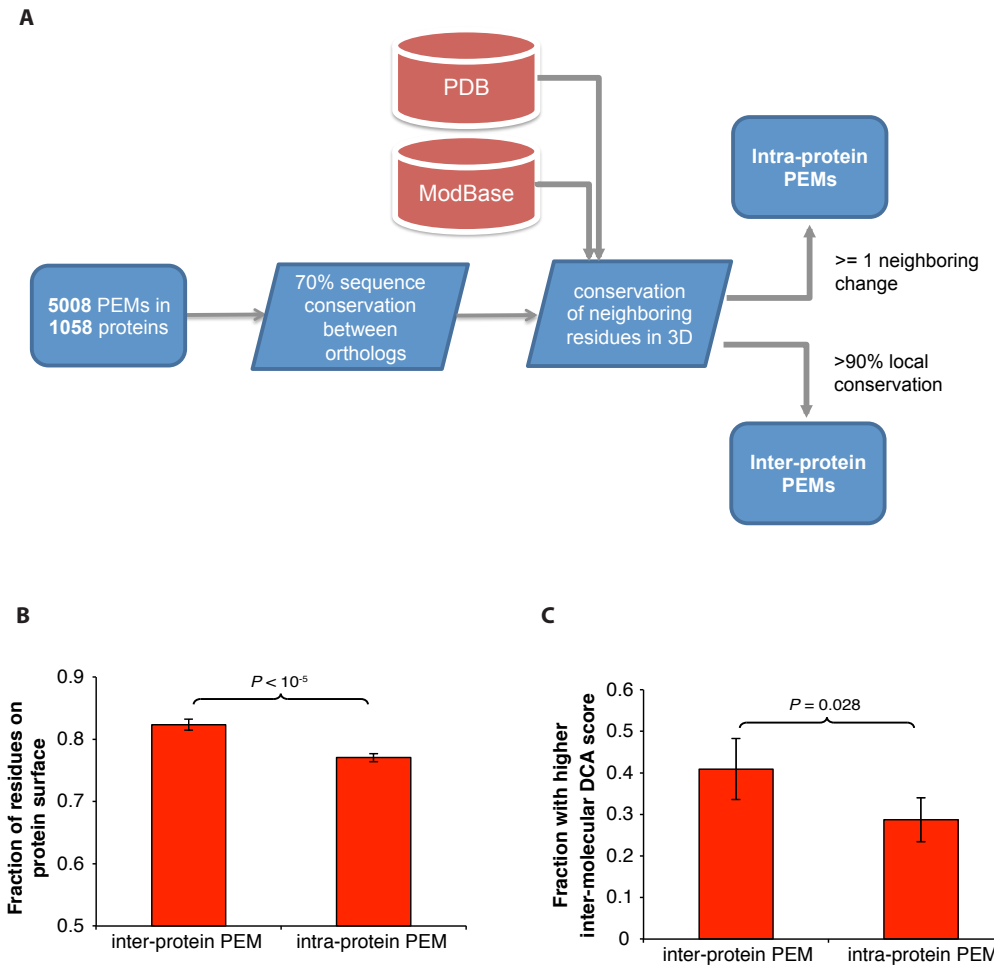


Figure 3.4 Identification of candidate inter-molecular and intra-molecular PEMs. **(A)** Pipeline to identify high-quality inter- and intra-molecular PEMs using 3D structures of proteins. **(B)** Fraction of surface residues among the candidate inter- and intra-molecular PEMs ($P < 10^{-5}$). **(C)** Fraction of PEMs with higher inter-molecular DCA score than intra-molecular DCA score ($P = 0.028$).

Currently, only about 40% of all proteins have a crystal structure or a high quality model. To expand the number of candidate inter-molecular PEMs for experimental evaluation, we used a linear local conservation cutoff instead of the 3D local conservation cutoff for proteins without available structure or model. Here,

instead of considering residues within 8Å of a PEM as its neighbors, we consider 10 amino acids upstream and downstream of a PEM as its neighbors. PEMs with local conservation rates above 90% are considered inter-molecular PEMs. Using the linear pipeline, we identified a total of 834 PEMs on 327 genes.

With the set of candidate PEMs identified, we investigated the impact of inter-molecular PEMs on protein-protein interactions using a system of high-throughput site-directed mutagenesis and Y2H[19]. Only PEMs with potential inter-molecular compensatory mutations were included in the Y2H analysis (see section 3.2.4). In total, we obtained the Y2H phenotypes of 113 interactions of 41 inter-molecular PEMs on 24 genes. This includes high-quality inter-molecular PEMs identified above, as well as a set of PEMs located on the protein interaction domains. We found that PEMs located on the interaction domains disrupts 33% of the interactions tested (5 out of 15 interactions disrupted). This is significantly lower than the 61% disruption rate of regular disease mutations located on the protein interaction domains[19] ($P = 0.028$ by cumulative binomial test; Figure 3.5). Similarly, for protein interactions without cocrystal structures (without structural information), only 12% (12 out of 98) of them are disrupted by PEMs, significantly lower than the 32% disruption rate of regular disease mutations without structural information ($P < 10^{-5}$ by cumulative binomial test; Figure 3.5). This is the first experimental evidence demonstrating that PEMs are less deleterious than regular disease mutations at the molecular level. Furthermore, the results also show that similar to regular disease mutations, PEMs at the protein interaction domains are more likely to interfere with protein-protein interactions.

Mutant clones for 31 intra-molecular PEMs were created using high-

throughput, site-directed mutagenesis[19]. Experimental candidates of intra-molecular PEMs were chosen based on the availability of ORFs in our lab and the presence of candidate intra-molecular compensatory mutations (see section 3.2.4). The impacts of the intra-molecular PEMs on protein stability will be tested with a GFP assay.

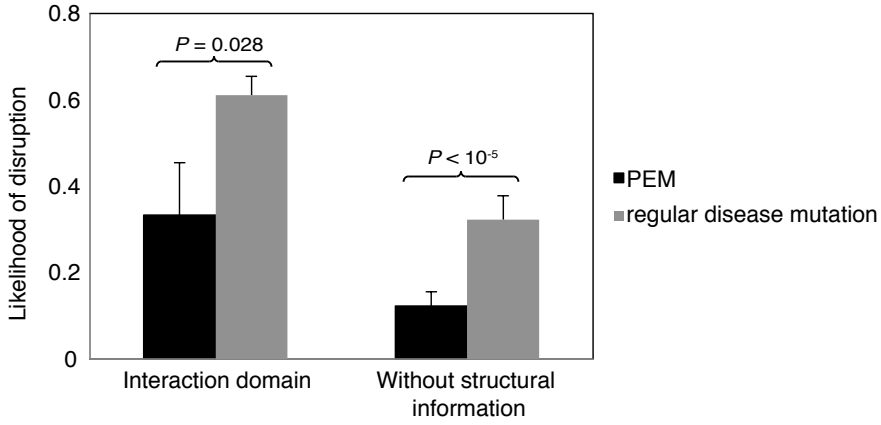


Figure 3.5 Effect of PEMs and regular disease mutations on protein-protein interactions. Likelihood of disruption of interactions by PEMs and regular disease mutations given they are on the protein interaction domain and in the absence structural information.

3.2.4. Identification of potential compensatory mutations

Next, we systematically searched for compensatory mutations for PEMs that disrupt protein-protein interactions or protein stability, with the goal of experimentally validating the epistatic interactions between PEMs and their compensatory mutations. For inter-molecular PEMs that disrupted specific protein-protein interactions, we searched for potential compensatory mutations on the corresponding interacting proteins. For each interacting protein, we compared the sequences of the human protein and its orthologs in species where the human disease residue is the wildtype. To ensure that the orthologous pair of interacting proteins is not too diverged between

human and the other species, we again considered only the orthologous pairs with sequence identity above 70%. All residues on the orthologous protein that are different from their corresponding residues in the human protein are considered potential compensatory mutations. On average, more than 2,000 potential compensatory mutations were found for each inter-molecular PEM. We used two criteria to evaluate the functional relevance of the potential compensatory mutations identified. First, compensatory residues should be functionally important sites that are conserved through evolution. Here we used the Jensen–Shannon divergence[20] as a measure of protein residue conservation (see Materials and Methods). Second, compensatory residues should coevolve with PEMs through evolution. To measure the extent of coevolution, we computed a correlation score that measures the co-occurrence of a PEM and a compensatory residue, and penalizes the co-occurrence of the human wildtype residue and the compensatory residue (see Materials and Methods). DCA is not used here as it is slow to compute, and therefore it is not feasible to compute DCA scores for large numbers of interacting proteins. In addition, DCA measures the coevolution of two protein sites, but not the correlation of specific amino acid residues at two sites through evolution. For each PEM, top compensatory mutation candidates with the highest correlation scores and with conservation scores above 0.4 are chosen for experimental testing. Their impacts on protein interaction with and without the corresponding PEM will be tested by Y2H.

For candidate intra-protein PEMs, we searched for potential compensatory mutations among all neighboring residues within 8Å of each PEM in the human protein. If a neighboring residue of a PEM in the human protein is different from the

corresponding residue in the orthologous protein of a species where the PEM is found as the wildtype, the corresponding residue in the other species is considered a potential compensatory mutation. To find the top compensatory mutation candidates, we estimated the stabilizing effects of each compensatory mutation on the free energy of the folded protein using Rosetta. For each PEM, we estimated the change in free energy upon mutation to the disease residue ($\Delta\Delta G_1 = \Delta G_{PEM} - \Delta G_{wildtype}$). Then we estimated the change in free energy after introducing the compensatory mutation ($\Delta\Delta G_2 = \Delta G_{PEM, CM} - \Delta G_{wild\ type}$). The stabilizing effect of the compensatory mutation is the difference between $\Delta\Delta G_2$ and $\Delta\Delta G_1$. The stabilizing effects of the top compensatory mutation candidates for each PEM will be evaluated experimentally using a GFP assay.

3.3. Discussion

In this study, we found that PEMs tend to be polymorphic in the human population and can exist at higher allele frequencies than regular disease mutations. We have also shown that PEMs have much milder effects on protein-protein interactions compared to regular disease mutations. This is the first experimental evidence demonstrating that PEMs are less disruptive than regular disease mutations at the molecular level. In terms of protein stability, based on $\Delta\Delta G$ predictions by Rosetta, Xu et al. (2014)[2] found that on average, PEMs only increases the folding energy of the proteins by 0.79 kcal/mol, significantly lower than the protein stability reduction by regular disease mutations ($\Delta\Delta G$ =2.25 kcal/mol on average). Given that most wildtype proteins have folding energies between -3 to -10 kcal/mol[21], it seems that most PEMs would have

very small impacts on protein stability. Paradoxically, it has been found that at the organismal level, PEMs are not enriched for associations with less severe or later onset diseases[4]. This raises the question, how do these PEMs lead to disease in human? One possibility is that PEMs can slightly lower the dissociation constants (Kd) of protein-protein interactions, and the small destabilizing effects of PEMs on the protein complexes could not be detected by our Y2H assay. However, small changes in the Kd of protein-protein interactions or the ΔΔG of proteins may still disrupt the equilibrium of the cell system, leading to disease. Given that the likelihood of disruption of protein-protein interactions by PEMs are similar to that of common SNPs (unpublished data), and that the ΔΔG caused by PEMs are only slightly higher than that cause by common SNPs (0.79 kcal/mol vs. 0.26 kcal/mol)[2], an alternative hypothesis is that many PEMs are benign by themselves, and will only lead to disease if there exist a synergistic mutation, which could be a common population variant, in the same person. More work need to be done to support the validity of the synergistic hypothesis.

In addition, we computationally predicted both inter-molecular and intra-molecular compensatory mutations for PEMs, and top candidates will be experimentally tested for molecular epistasis with PEMs. This is the first system-wide effort to identify epistatic partners of PEMs. The identification of compensatory mutations not only allow us to have a better understanding of the molecular interactions during disease pathogenesis, but also allow the design of compensatory drugs that can rescue the impacts of these disease mutations in human.

3.4. Materials and Methods

3.4.1. Protein structures and models

Protein crystal structures were downloaded from PDB and protein structural models were downloaded from ModBase. To ensure the confidence of our local conservation and ddG calculations, only PDB structures with at least 250 amino acids or cover at least 40% of the target Swiss-Prot protein sequence were retained. Also, only ModBase models with MPQS $\geq$ 1.1 were used, because models with MPQS $\geq$ 1.1 are considered to be reliable by ModBase. Finally, amino acid indices of all structures and models were mapped to the corresponding indices of the Swiss-Prot protein by pairwise realignment using SIFTS[22].

3.4.2. Identification of surface, interface and contact residues

An amino acid residue with more than 15% of its total surface area exposed to the solvent (a water molecule in this case) is defined as a surface residue. An interface residue is defined as a surface residue on a protein chain that undergoes at least 1Å^2 change in solvent accessible surface area upon complex formation. Residue pairs on interacting proteins whose alpha carbons are within 8Å of each other are defined as contact residues. All solvent accessibility calculations were done using NACCESS[23].

Interacting domains were identified as described in Das et al. (2014)[15]. In short, we first identified a set of putative interacting domains using the "homology modeling" approach described earlier[24]. To reduce the false positive rate of the homology modeling approach, we kept only putative interacting domains that contain at least one interface residue. In addition, to identify true interacting domains missed

by the homology modeling approach, we included protein domains not identified by the homology modeling method if they contain 5 or more interface residues.

3.4.3. Identification of neighboring residues

Residues are defined as neighbors of a PEM if their alpha carbons are within 8Å to the alpha carbon of the PEM. The alpha carbon distances are calculated based on the atomic coordinates given in the PDB file.

3.4.4. Coevolution and conservation scores calculation

DCA scores are calculated using the Matlab implementation of DCA[16]. DCA is used as the measure of coevolution here because it is able to distinguish between direct correlations between protein sites from indirect correlations. Inter-protein DCA scores were calculated by concatenating the protein sequences of the two interacting partners in each species where both orthologs are present. DCA scores are only calculated for proteins with at least 50 orthologs.

The conservation of protein sites across the 100 vertebrate species is estimated using the Jensen–Shannon divergence[20]. The Jensen–Shannon divergence is preferred over the commonly used relative entropy to measure protein conservation because it is symmetric and is bounded from zero to one. The Jensen–Shannon divergence scores are calculated using a Python implementation by Capra and Singh (2007)[25].

Correlation scores are calculated as follows:

$$\textbf{Correlation score} = \varphi(WT_{PEM}, WT_{CM}) + \varphi(MUT_{PEM}, MUT_{CM}) - \varphi(WT_{PEM}, MUT_{CM}) - \varphi(MUT_{PEM}, WT_{CM})$$

WT_{PEM}, WT_{CM}, MUT_{PEM}, MUT_{CM} are Boolean vectors of length n, where n is the number of species. Each entry in WT_{PEM} describes whether the amino acid residue at the PEM site of a species is the human wildtype residue. Each entry in MUT_{PEM} describes whether the amino acid residue at the PEM site of a species is the human disease residue. Each entry in WT_{CM} describes whether the amino acid residue at the potential compensatory site of a species is the human wildtype residue. Each entry in MUT_{CM} describes whether the amino acid residue at the potential compensatory site of a species is the compensatory residue. φ is the phi coefficient of the two vectors. It measures the degree of association between two binary variables.

3.4.5. ΔΔG Predictions using Rosetta

First, protein structures were preminimized for folding energy using the "ddg_min" program in Rosetta with default parameters. Then the changes in folding energy upon point mutations are estimated using the low-resolution protocol of the "ddg_monomer" program with default parameters.

3.4.6. Construction of plasmids and PEM mutant clones

Mutant clones were generated by site-directed mutagenesis as describe in Wei et al. (2014)[19].

3.3.7. Yeast two-hybrid

Y2H assays were performed as describe in Wei et al. (2014)[19].

REFERENCES

1 Kondrashov, A. S., Sunyaev, S. & Kondrashov, F. A. Dobzhansky-Muller incompatibilities in protein evolution. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 14878-14883, doi:10.1073/pnas.232565499 (2002).

2 Xu, J. & Zhang, J. Why human disease-associated residues appear as the wild-type in other species: genome-scale structural evidence for the compensation hypothesis. *Molecular biology and evolution* **31**, 1787-1792, doi:10.1093/molbev/msu130 (2014).

3 Ferrer-Costa, C., Orozco, M. & de la Cruz, X. Characterization of compensated mutations in terms of structural and physico-chemical properties. *Journal of molecular biology* **365**, 249-256, doi:10.1016/j.jmb.2006.09.053 (2007).

4 Gao, L. & Zhang, J. Why are some human disease-associated mutations fixed in mice? *Trends in genetics : TIG* **19**, 678-681, doi:10.1016/j.tig.2003.10.002 (2003).

5 Poon, A., Davis, B. H. & Chao, L. The coupon collector and the suppressor mutation: estimating the number of compensatory mutations by maximum likelihood. *Genetics* **170**, 1323-1332, doi:10.1534/genetics.104.037259 (2005).

6 Baresic, A. & Martin, A. C. Compensated pathogenic deviations. *Biomolecular concepts* **2**, 281-292, doi:10.1515/bmc.2011.025 (2011).

7 Baresic, A., Hopcroft, L. E., Rogers, H. H., Hurst, J. M. & Martin, A. C. Compensated pathogenic deviations: analysis of structural effects. *Journal of molecular biology* **396**, 19-30, doi:10.1016/j.jmb.2009.11.002 (2010).

8 Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Human mutation* **21**, 577-581, doi:10.1002/humu.10212 (2003).

9 Stenson, P. D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome medicine* **1**, 13, doi:10.1186/gm13 (2009).

10 Miller, W. *et al.* 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome research* **17**, 1797-1808, doi:10.1101/gr.6761107 (2007).

11 Rosenbloom, K. R. *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic acids research* **43**, D670-681, doi:10.1093/nar/gku1177 (2015).

12 Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-69, doi:10.1126/science.1219240 (2012).

13 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-249, doi:10.1038/nmeth0410-248 (2010).

14 Wang, X. *et al.* Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature biotechnology* **30**, 159-164, doi:10.1038/nbt.2106 (2012).

15 Das, J. *et al.* Elucidating common structural features of human pathogenic variations using large-scale atomic-resolution protein networks. *Human mutation* **35**, 585-593, doi:10.1002/humu.22534 (2014).

16 Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America* **108**, E1293-1301, doi:10.1073/pnas.1111471108 (2011).
17 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic acids research* **28**, 235-242 (2000).
18 Pieper, U. *et al.* ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic acids research* **42**, D336-346, doi:10.1093/nar/gkt1144 (2014).
19 Wei, X. *et al.* A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. *PLoS genetics* **10**, e1004819, doi:10.1371/journal.pgen.1004819 (2014).
20 Lin, J. H. Divergence Measures Based on the Shannon Entropy. *Ieee T Inform Theory* **37**, 145-151, doi:Doi 10.1109/18.61115 (1991).
21 Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Current opinion in structural biology* **19**, 596-604, doi:10.1016/j.sbi.2009.08.003 (2009).
22 Velankar, S. *et al.* SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic acids research* **41**, D483-489, doi:10.1093/nar/gks1258 (2013).
23 Hubbard, S. & Thornton, J. 'NACCESS', computer program. (1993).
24 Meyer, M. J., Das, J., Wang, X. & Yu, H. INstruct: a database of high quality three-dimensional structurally resolved protein interactome networks. *Bioinformatics*, doi:10.1093/bioinformatics/btt181 (2013).
25 Capra, J. A. & Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875-1882, doi:Doi 10.1093/Bioinformatics/Btm270 (2007).

CHAPTER 4

IDENTIFICATION AND CHARACTERIZATION OF NOVEL DISEASE PREDISPOSING VARIANTS IN HUMAN HEREDITARY DISEASES

The work in this chapter is done in collaboration with Dr. Steven Lipkin at the Weill Cornell Medical College, with Dr. Xiaomu Wei as the leading scientist. Dr. Jian Sun performed the sequence alignment and variant calling. Dr. Xiaomu Wei performed Sanger sequencing, TaqMan assays and protein stability assays. Dr. Xiaomu Wei also contributed to the variant filtering pipelines as well as the curation of biologically relevant genes in Crohn disease and Multiple Myeloma. Dr. Manirath Khounlotham and Dr. Melissa Rose performed the Flow Cytometry analysis. The statistical analysis in the mutation burden test is performed in collaboration of David Sinclair and Dr. James Booth.

4.1. Introduction

Exploring the complex genotype-phenotype relationships has been a major focus of genetic studies. Identification of genes and mutations that are involved in pathogenesis not only provides insights into the biology of the diseases, but also has great implications in the clinical diagnostics and treatment of diseases. Traditionally, disease genes and mutations were often identified through linkage mapping and candidate gene resequencing. However, the genetic bases of many Mendelian diseases have yet to be identified. In the Online Mendelian Inheritance in Man (OMIM) database, there are 7,899 phenotypes with suspected Mendelian basis, however, only 4,405 phenotypes have known molecular basis[1]. Genome-wide approaches of risk-gene identification are needed to bridge the gap of knowledge between human diseases and their underlying genetics. Genome-wide association studies (GWAS) have found a large number of associations between common variants (minor allele frequency >5%) in the human genome and common human diseases. However, the common variants genotyped in GWAS are meant to tag the genomic loci that contribute to disease, and in most cases the actual causal gene/mutation could not be identified. In addition, in most GWA studies, the genetic loci identified generally have very small effect sizes, and together only explain a small fraction of the observed heritability of the phenotype[2,3]. Furthermore, it is believed that many human diseases could be attributed to rare variants, which are believed to be more deleterious[4].

With the decreasing sequencing cost, it is now feasible to sequence the entire genome or exome of large numbers of diseased and control individuals, and identify almost all genetic variants in these individuals. Thus, whole exome/genome

sequencing provides a powerful platform for the identification of causal genes and mutations that underlie Mendelian and complex diseases in humans[5,6]. However, the identification of causal variants among large numbers of benign polymorphisms and sequencing errors remains a major challenge[7].

Here, we investigated two familial inherited diseases, Crohn disease and multiple myeloma. We performed whole-exome and -genome sequencing on patients and families with inherited Crohn disease and multiple myeloma. By setting up a novel variant prioritizing pipeline that incorporates various biological data such as known disease genes, biological pathways, protein-protein interactions and protein structures, we aim to identify predisposing genetic factors leading to each disease.

4.2. Familial Crohn Disease

4.2.1. Study design

Inflammatory bowel disease (IBD) is an autoimmune illness characterized by chronic inflammation of the gastrointestinal tract, and it affects approximately 1.4 million Americans. The two most common forms of IBD are Crohn disease, characterized by patchy transmural inflammation from mouth to anus and ulcerative colitis, characterized by continuous mucosal inflammation of the colon[8]. Both genetic and environmental factors contribute to the development of IBD. It has been proposed that the host-microbe interactions in the gut play an important role in the pathogenesis of IBD, where intestinal microbes elicit exaggerated inflammatory responses in genetically susceptible hosts[9,10]. It has been observed that IBD often affect multiple members of the same family, suggesting a substantial genetic component in IBD,

especially in Crohn disease[11,12]. Genome-wide association studies have identified over 160 risk loci associated with IBD[10]. However, most of GWAS variants only confer low risk to the disease and are therefore not clinically actionable. Here, by performing whole-genome and whole-exome sequencing on clinically well-characterized pediatric Crohn disease kindreds (patient samples were collected by Dr. Melissa Rose at the New York Presbyterian Hospital), we aim to identify the first high penetrance, familial inherited pediatric Crohn disease risk genes and variants. We performed whole-genome sequencing on 10 individuals of 3 pediatric Crohn disease families, and whole-exome sequencing on 12 individuals of 2 pediatric Crohn disease families (Figure 4.1). By performing Mendelian segregation analysis and a novel variant prioritizing pipeline, we identified interesting cosegregating variants in these families and functionally validated one of them. The identification of Crohn disease high-risk genes promotes the development of new genetic diagnostic tests as well as the design of new mechanistically based targeted drugs, to allow early detection of Crohn disease and improved treatment for patients and their at-risk family members.

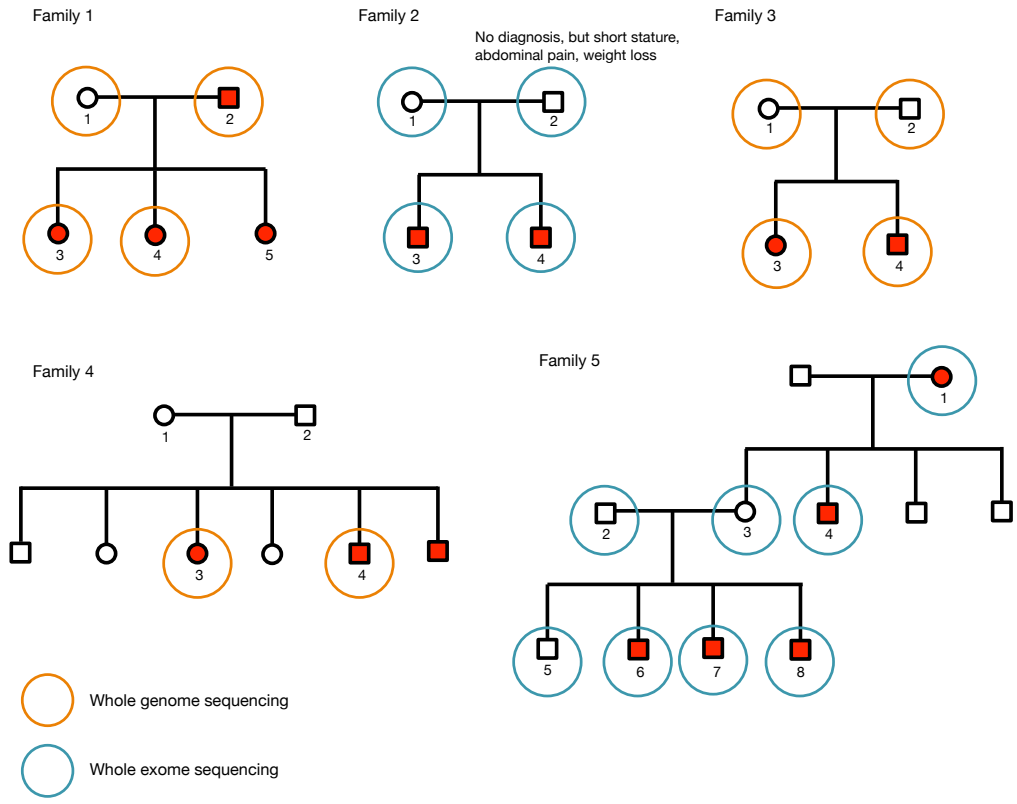


Figure 4.1 Pedigrees of the five pediatric Crohn disease families sequenced. Individuals with Crohn disease are represented by filled red symbols. Individuals that are not diagnosed with Crohn disease are represented by unfilled symbols. Individuals with whole-genome or whole-exome sequencing are circled with orange or blue circles respectively.

4.2.2. Mendelian segregating analysis

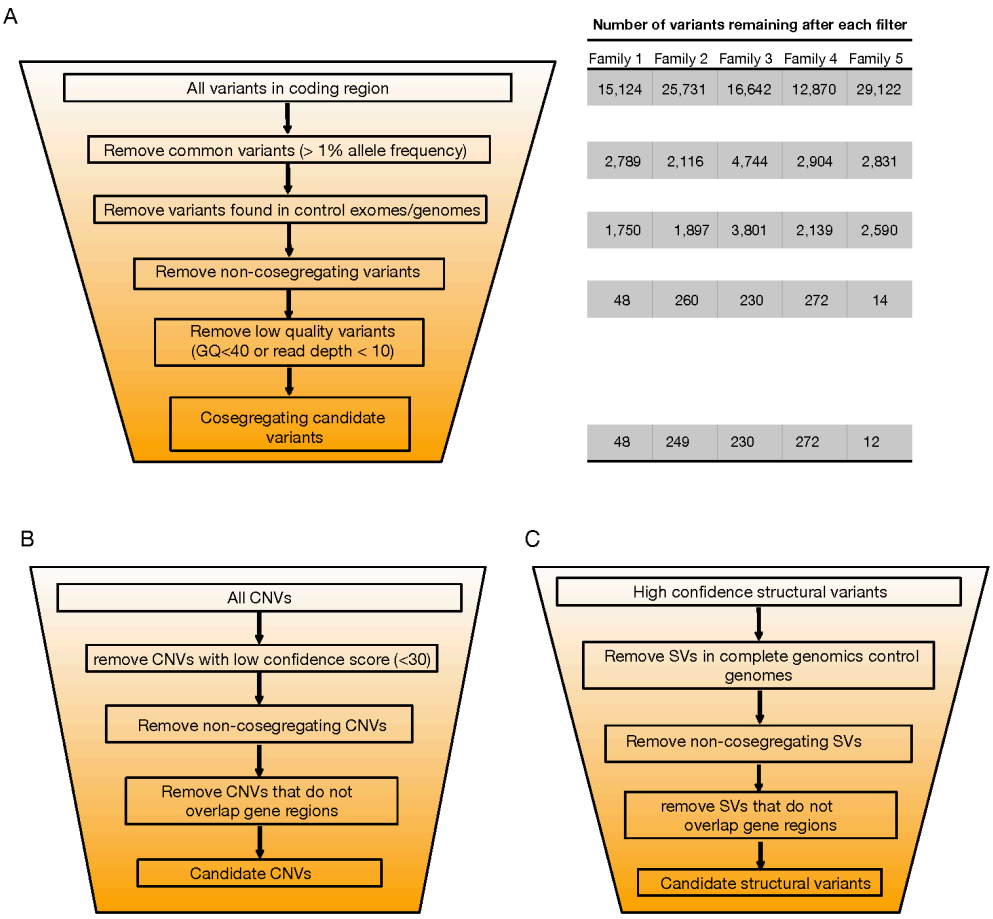
The whole-genome sequencing was performed by Complete Genomics. Each genome was sequenced to a mean depth of 40x. Sequence alignment, variant calling, copy number variant and structural variant discovery were performed and delivered by Complete Genomics. For the whole-exome sequencing, exome capture was performed using Agilent SureSelect Human All Exon 50MB paired-end capture, each sample was barcoded and sequenced on Illumina HiSeq2000's to a mean depth of 100x. The high

sequencing depth increases the accuracy and sensitivity of discovering novel heterozygotes. Reads were aligned to human genome build GRCh37 using BWA[13]. Following the GATK Best Practices guideline, we performed duplicate reads removal, base quality score recalibration, and dynamic realignment at suspected indels, following by SNP and indel discovery[14-16].

Causal mutations are likely to be functional mutations that alter the sequence and structure of the encoded protein. Therefore, we first focused our analyses on protein sequence altering variants in the coding region, including missense, nonsense, frameshift, and splice site variants. There are 15,000 to 30,000 of such variants found in each of the 5 families we sequenced. Next, we discarded common polymorphisms (>1% allele frequency) in 1000 Genomes[17] and NHLBI GO Exome Sequencing Project (ESP) data[18], as high-penetrance Crohn disease predisposing variants should be rare in the general population. Then, we used sequencing platform matched controls to remove platform specific sequencing errors in our samples. For the 3 families sequenced by Complete Genomics, the control set used is the Complete Genomics public whole-genome sequencing data of 54 unrelated, non-diseased individuals. For the 2 families with whole-exome sequencing, the control set used include 17 multiple myeloma patients that were sequenced on the same platform. All variants that were also found in the matched control set were discarded. In each individual, we only kept high quality variants with genotype quality score of at least 40 and read depth of at least 10. To identify candidate predisposing variants, only variants that co-segregate with the disease phenotype will be retained for further analyses. From the family pedigrees, we assumed dominant inheritance in families 1, 2

and 5, where we retained variants that are found in all affected family members and are not found in the unaffected parent. Note that in family 2, although the father was not diagnosed with Crohn disease, he showed disease symptoms such as short stature, abdominal pain, and weight loss. Therefore we consider him as diseased in the Mendelian segregation analysis. In families 3 and 4, it is unclear whether the inherited genetic factor is recessive or partially penetrant dominant. Therefore, in the Mendelian segregation analysis, we kept all variants shared between the affected siblings. Using this variant filtering pipeline, we found 12 to 272 high quality, cosegregating variants in each family (Figure 4.2A).

In families with whole-genome sequencing, we also analyzed their copy number variants (CNVs) and structural variants (SVs) in addition to small variants (Figure 4.2B-C). Both CNVs and SVs were called and annotated by the Complete Genomics standard data processing pipeline. We did not find any novel, high-quality cosegregating CNV that overlap gene regions. We found a total of 5 high-quality cosegregating SVs that overlap gene regions in the 3 families analyzed (Table 4.1). Four of these SVs only affect the intronic regions of the genes they overlap. A cosegregating deletion found in family 4 deletes exons 4 and 5 of the growth hormone receptor (*GHR*) gene. *GHR* is a known disease gene that has been associated with Laron dwarfism, short stature and hypercholesterolemia, but currently it is not known to be related to Crohn disease. More genetic and functional validations are needed to determine if the deletion in *GHR* is involved in Crohn disease in family 4.



Number of variants remaining after each filter

Family 1	Family 2	Family 3	Family 4	Family 5
15,124	25,731	16,642	12,870	29,122
2,789	2,116	4,744	2,904	2,831
1,750	1,897	3,801	2,139	2,590
48	260	230	272	14
48	249	230	272	12

Figure 4.2 (A) The variant filtering pipeline for SNVs and small indels. Table on the right shows the number of variants left after each filter in each family. Variant filtering pipelines for **(B)** copy number variants, and **(C)** structural variants.

Family	Chromosome region	Type	Overlapping genes	Affect coding region?
CD1	chr7: 3654449 - 3655539	deletion	**SDK1**	intron
CD3	chr2: 72440497 - 72441362	distal-duplication	**EXOC6B**	intron
	chr14: 63369974 - 63379358	deletion	**KCNH5**	intron
CD4	chr1: 72015921 - 72054939	deletion	**NEGR1**	intron
	chr5: 42693088 - 42701005	deletion	**GHR**	Deletes exons 4 and 5
	chr10: 12867123 – 12867190	distal-duplication	**CAMK1D**	intron

Table 4.1 Cosegregating structural variants that overlap genes in families 1, 3 and 4.

4.2.3. Prioritization of candidate variants using protein interaction network and biological pathways

After the Mendelian segregation analysis and variant filtering, there are 12 to 272 small variants left in each of the families. To further narrow down the candidate variant list for further genetic and functional validations, we focused on a set of 754 genes that are relevant to the biology of Crohn disease. This set of biological relevant genes includes known disease genes associated with Crohn disease, IBD or ulcerative colitis curated from the Human Gene Mutation Database (HGMD)[19,20] and OMIM[1,21] and their interacting partners in the protein-protein interaction network[22], a list of the most clinically relevant and highly significant Crohn disease genes found by GWAS[23] and their interaction partners, and all genes in innate immune response related or autophagy related Gene Ontology terms (Figure 4.3). Genes involved innate immune response and autophagy are included as innate immune response and autophagy are essential in maintaining host-microbe balance in the gut, and multiple the innate

immune genes and autophagy genes have been implicated in Crohn disease[9,23].

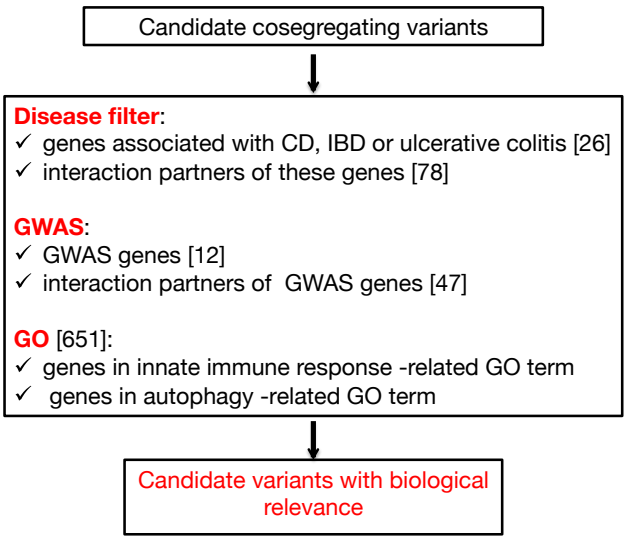


Figure 4.3 Identification of variants on genes that are biologically relevant to Crohn disease. Modified from a figure made by Dr. Xiaomu Wei.

We found a total of 18 candidate variants on biologically relevant genes in the 5 Crohn disease families sequenced. Interestingly, among the six candidate variants from family 2 and family 5, four affect components of the Toll-like receptor 4 (TLR4) signaling pathway. TLR4 signaling is important in the induction of both innate and adaptive immune responses upon pathogen recognition[24]. Excessive microbe-induced immune response can lead to inappropriate inflammation, as observed in autoimmune diseases like IBD and rheumatoid arthritis. We hypothesize that our variants the TLR4 pathway genes increase the activity of TLR4 and NOD2 signaling, leading to increased production of proinflammatory cytokines and inflammation in the guts.

We used Polyphen2 and SIFT to predict the potential impacts of the 6 candidate variants, and 5 of them are predicted to be deleterious by at least one

algorithm. Missense mutations can disrupt protein functions by disrupting specific protein-protein interactions or destabilizing the whole protein. To check if our candidate variants can potentially affect protein stability, we used Rosetta[25] to predict the change in protein folding energy brought about by each mutation ($\Delta\Delta G=\Delta G_{mut}-\Delta G_{wt}$). Mutations with more positive $\Delta\Delta G$ are predicted to be more destabilizing. We found that one of our candidate variants segregating in family 2 has a large $\Delta\Delta G$ of 4.003. We hypothesize that this mutation is a loss-of-function mutation that destabilizes the protein, and thereby abolish the regulatory effect of the protein on TLR4 signaling, leading to increased cytokine production and inflammation. This potential loss-of-function variant is prioritized for further genetic and functional validations.

4.2.4. Genetic and functional validations of candidate variant

Our candidate loss-of-function variant was first validated in the affected individuals by Sanger sequencing. A disease-causing gene is more clinically relevant if a large fraction of diseased individuals is affected by mutations on the gene. To study if our candidate variant is found in other Crohn disease patients, we performed Taqman assay for the candidate variant on a panel of 390 Crohn disease samples and 437 control samples obtained from the Crohn's and Colitis Foundation of America (CCFA). We found our candidate loss-of-function variant in another Crohn disease patient and not in any controls. This is highly interesting as this mutation might be a causal variant in another Crohn disease patient, and is not private to the family we sequenced.

Next, protein stability assays confirms our hypothesis that our candidate variant is a loss-of-function variant that destabilizes the protein. Furthermore, using flow cytometry, we have shown that the candidate variant decreases the B cell surface expression of the protein in affected individuals of family 2, confirming the effect of the candidate mutation in the patients.

4.2.5. Conclusions

In this study, we performed whole genome/exome sequencing on 22 individuals from 5 pediatric Crohn disease families. Using Mendelian segregation analysis and a novel variant prioritizing pipeline that incorporates information on known Crohn disease risk genes, protein-protein interaction network, biological pathways, and protein structures, we identified a potential high-penetrance risk variant associated with Crohn disease. We have shown that the loss-of-function variant destabilizes the protein. Without the negative regulation of the candidate protein, the TLR4 signaling pathway is over active, causing inappropriate inflammation in patients with the candidate mutation. More functional studies will be done to confirm the effect of the candidate mutation on downstream cytokine production.

4.3. Familial and Early Onset Multiple Myeloma

4.3.1. Study design

Multiple Myeloma is a type of hematological malignancy characterized by the excessive proliferation of monoclonal plasma cells. It is the second most common form of hematological cancer and accounts for about 13% of all hematological cancers[26]. Multiple myeloma commonly progresses from an asymptomatic precursor disease, monoclonal gammopathy of undetermined clinical significance (MGUS)[27]. With additional chromosomal translocations or genetic mutations, MGUS can evolve into symptomatic myeloma, characterized by anemia, renal insufficiency and lytic bone lesions[26-28]. Multiple myeloma is currently incurable and has a 10-year survival rate of about 30%[26]. Familial clustering of multiple myeloma and MGUS has been observed[29], and the relative risk of multiple myeloma in first-degree relatives of patients was estimated to be 2-5%[30,31]. The inheritance patterns of multiple myeloma in these families suggest the presence of highly penetrant, autosomal dominant risk genes[32].

In this study, we aim to identify predisposing germline genetic variants that increase the risk of multiple myeloma. We performed whole-exome sequencing on the constitutional DNA of 113 multiple myeloma patients. These include early onset multiple myeloma patients (age of onset lower than 45 years old) and probands with at least one first-degree relative affected by myeloma or associated B-cell malignancies and. Multiple myeloma is a late-onset illness affecting primarily people above 65 years old. Only about 2% of the patients are younger than 40 years old[33]. Therefore, it is likely that patients with early onset multiple myeloma harbor predisposing genetic

factors that decreased the age of onset of the disease. In addition, we also sequenced multiple family members of 7 kindreds with familial multiple myeloma (19 affected individuals in total). Patient samples were collected by Dr. Henry Lynch, Dr. Carrie Snyder and Dr. Dina Becirovic at Creighton University, Dr. Nicola Camp at University of Utah, Dr. Kenneth Offit at Memorial Sloan Kettering Cancer Center, Dr. Ruben Niesviesky and Dr. David Jaybaylan at Weill Cornell Medical College, and our collaborators at University of Michigan.

4.3.2. Identification of recurrent variants in multiple myeloma

Exome capture was performed using Agilent SureSelect Human All Exon 50MB paired-end capture, each sample was barcoded and sequenced on Illumina HiSeq2000's to a mean depth of 100x. Reads were aligned to human genome build GRCh37 using BWA[13]. Following the GATK Best Practices guideline, we performed duplicate reads removal, base quality score recalibration, and dynamic realignment at suspected indels, following by SNP and indel discovery[14-16].

We aim to identify high-risk multiple myeloma variants that are rare in the general population but occur recurrently in individuals with multiple myeloma. We focused our recurrent variant analyses on coding variants that alter protein sequences, including missense, nonsense, frameshift, and splice site variants. To retain only high quality variants that are rare in the general population, we first discarded common polymorphisms (>0.5% allele frequency) found in 1000 Genomes[17], ESP[18], and the Exome Aggregation Consortium (ExAC) data in the corresponding populations. Then, to remove sequencing platform specific biases, we compared variants found in

multiple myeloma exomes against a set of 45 sequencing platform matched control exomes, comprising 10 lung cancer constitutional exomes, 15 breast cancer constitutional exomes and 20 Crohn disease exomes. Only multiple myeloma specific variants that are not found in the control exomes were retained. In each individual, we kept only high quality variants with genotype quality score of at least 40 and read depth of at least 10. For families with more than 1 affected individuals sequenced, we only considered consegregating variants in each family. In total, there are 5,790 rare, coding variants that occur in 2 or more individuals (singletons) or families.

4.3.3. Prioritization of candidate variants using protein interaction network and biological pathways

To further filter the candidate variant list for multiple myeloma predisposing variants, we compiled a set of 4,261 genes that are biologically relevant to multiple myeloma, and focused further efforts on variants on these genes. This set of biological relevant genes includes known disease genes associated with multiple myeloma curated from HGMD[19,20], OMIM[1,21] and Malacards[34]; protein interaction partners of multiple myeloma disease genes[22], genes in the same pathway as multiple myeloma disease genes curated from Biocarta, KEGG[35] and Reactome[36,37]; GWAS genes associated with multiple myeloma curated from the NHGRI GWAS Catalogue[38]; somatically mutated multiple myeloma genes[39]; as well as genes in the NFKB pathway, as they are frequently somatically mutated in multiple myeloma[39] (Figure 4.4). Out of the 5,790 recurrent variants, 900 variants were found on biologically relevant genes and are kept for further analysis.

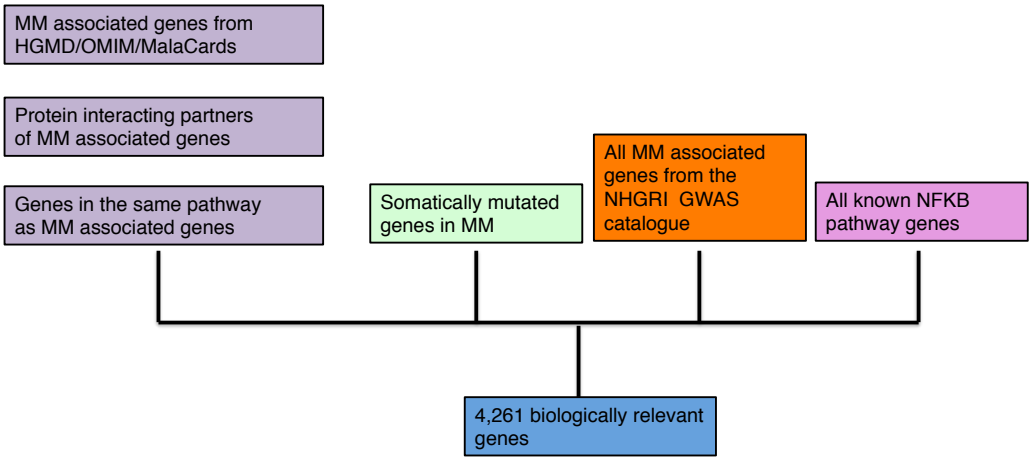


Figure 4.4 Compiling a set of genes that are biologically relevant to multiple myeloma (MM). Modified from a figure made by Dr. Xiaomu Wei.

4.3.4. Identification of highly mutated genes by mutation burden test

Although there could be specific predisposing mutations that occur recurrently in multiple diseased individuals, there could also be risk genes that harbor many different mutations in different multiple myeloma patients. Here, we performed a mutation burden test to identify such highly mutated genes in multiple myeloma.

With the increasing number of disease genome/exome sequencing studies, many tests have been developed to identify genes that are more highly mutated in the diseased individuals compared to random expectations. Methods such as CaMP[40], MuSiC[41] and MutsigCV[42] have successfully identified significantly mutated genes in somatic tumors by estimating gene-specific background mutation rates from synonymous mutations counts. However, the background mutation rate estimation may not be accurate for germline sequencing studies as germline mutations accumulate over generations and silent and non-silent mutations may have very different selection pressure. Indeed, I did not find any significantly mutated genes

using MutsigCV on the set of rare variants identified from our multiple myeloma exomes. For case-control sequencing studies, which are mostly studies on neurological disorders, rare variant association tests, also called mutation burden tests, were used to identified genes with excessive rare alleles in diseased individuals compared to controls. Rare variant association test is similar in concept to genome wide association studies, but since rare variants occur too infrequently, there is not enough power to perform association testing on individual locus. Instead, rare variants need to be aggregated into sets (e.g. aggregate by genes), and the mutation frequency of each set is compared between cases and controls[43,44].

In this study, we did not sequence healthy controls for comparison with multiple myeloma patients. Instead, we used the high-coverage exome sequencing results from the 1000 Genomes Project as controls in our mutation burden test. We compared the rare variant counts in each gene between 104 multiple myeloma patients of European ancestry and 298 1000 Genomes controls of non-Southern European ancestry. The procedure used to identify high-quality rare variants in multiple myeloma and control exomes is summarized in Figure 4.5. For the multiple myeloma cohort, we kept only multiple myeloma specific rare variants by removing variants found in our Crohn disease, lung cancer or breast cancer exomes. In addition, we also removed positions with low sequencing coverage in current public exome sequencing data, as the allele frequencies at these positions are of low confidence. From our previous Sanger sequencing validations on a small number of candidate variants, we found that variants that are present in more than 4 multiple myeloma samples but are rare in public databases are most likely sequencing errors. To remove these platform-

specific sequencing biases from our final mutation counts, we imposed a 2% intra-cohort allele frequency cutoff for both the multiple myeloma samples and the 1000 Genomes controls. Within either the case or control cohort, rare alleles (allele frequency < 0.5% in ESP) that are present in greater than 2% allele frequency were removed.

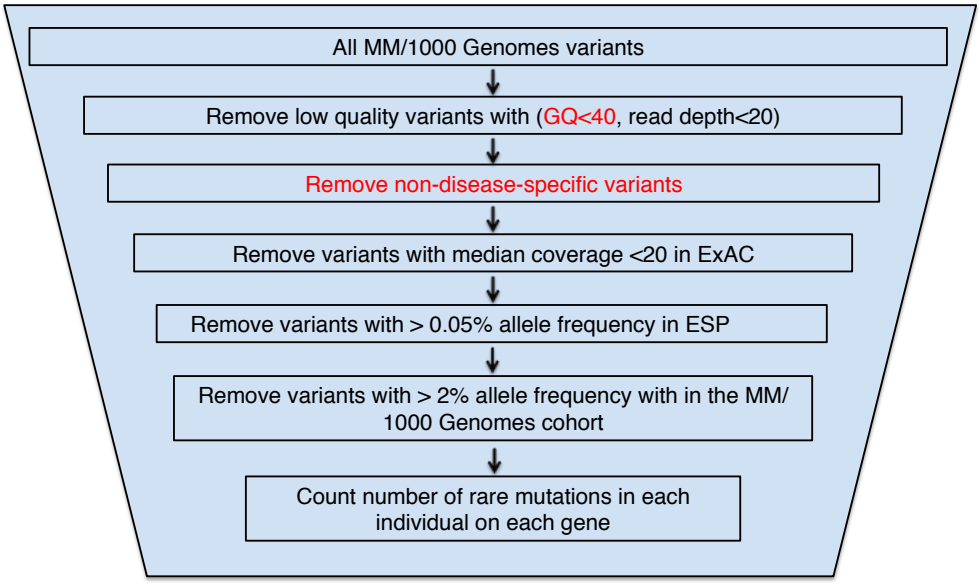


Figure 4.5 Generation of rare variant lists for mutation burden test from multiple myeloma (MM) and 1000 Genomes exome sequencing data. Filters in red were only applied to MM exomes.

One problem with using an external dataset is that the 1000 Genomes samples and our samples were sequenced at different times on different platforms. To control for possible batch bias, we calculated the coverage for each gene in each sample and normalized our sample mutation counts with gene coverage. Also, since we used an additional filter in the multiple myeloma samples by considering only disease-specific variants, the mutation counts in multiple myeloma samples tend to be lower than that in the control samples. To account for that, we multiplied the mutations counts in

controls by a constant factor such that the overall mutation rate in diseased and control individuals are the same. Then using a Poisson generalized linear model, we test whether there is a fold difference in mutation rate between multiple myeloma and control individuals for each gene. We performed a one-sided Z-test to determine the significance level of each gene.

$$H_0 : \frac{Y_{MM}/b_{MM}}{Y_c/b_c} = 1 \quad \text{vs} \quad H_a : \frac{Y_{MM}/b_{MM}}{Y_c/b_c} > 1$$

Where Y_{MM} and Y_c are the total number of mutation over all samples in the multiple myeloma group and the control group respectively, and b_{MM} and b_c are the total number of covered bases in the multiple myeloma group and the control group respectively.

In total, we found 120 genes that have significantly higher number of rare variants in the multiple myeloma group than in the control group after Benjamini–Hochberg false discovery rate correction. Among them, 27.5% (33 genes) are biologically relevant to multiple myeloma. Among all genes tested, only 21.9% are biologically relevant to multiple myeloma. Our list of significantly mutated genes seems to be enriched in biologically relevant genes, but this enrichment is not significant (P=0.08 by a one-sided Chi-squared test)

4.3.5. Genetic validation of candidate genes and variants

The top candidates from the recurrent mutation analysis, the mutation burden test, and the list of cosegregating mutations in the 7 multiple myeloma families were manually curated by Dr. Xiaomu Wei to generate a candidate list of genes for further genetics

validation. Each candidate variant was examined in the Integrative Genome Viewer[45] to remove variant call false positives. Also, since about 20 multiple myeloma patients we sequenced are of Ashkenazi Jewish descent, we compared our variant list to the variant data of 128 Ashkenazi Jewish genomes sequenced by The Ashkenazi Genome Consortium (TAGC)[46] to make sure that our candidates are not Ashkenazi Jewish founder mutations. Eventually, a final list of 20-30 candidate genes will be generated, and we will perform full-length targeted sequencing on these genes in about 1000 multiple myeloma patient samples and 1000 control sample. Variants that occur with a higher frequency on multiple myeloma samples than in controls will be further studied with functional assays.

REFERENCES

1 Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic acids research* **37**, D793-796, doi:10.1093/nar/gkn665 (2009).

2 Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18-21, doi:10.1038/456018a (2008).

3 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753, doi:10.1038/nature08494 (2009).

4 Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nature genetics* **40**, 695-701, doi:10.1038/ng.f.136 (2008).

5 Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nature reviews. Genetics* **12**, 745-755, doi:10.1038/nrg3031 (2011).

6 Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature reviews. Genetics* **11**, 415-425, doi:10.1038/nrg2779 (2010).

7 Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature reviews. Genetics* **12**, 628-640, doi:10.1038/nrg3046 (2011).

8 Bousvaros, A. *et al.* Differentiating ulcerative colitis from Crohn disease in children and young adults: report of a working group of the North American Society for Pediatric Gastroenterology, Hepatology, and Nutrition and the Crohn's and Colitis Foundation of America. *Journal of pediatric gastroenterology and nutrition* **44**, 653-674, doi:10.1097/MPG.0b013e31805563f3 (2007).

9 Abraham, C. & Cho, J. H. Inflammatory bowel disease. *The New England journal of medicine* **361**, 2066-2078, doi:10.1056/NEJMra0804647 (2009).

10 Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-124, doi:10.1038/nature11582 (2012).

11 Orholm, M., Fonager, K. & Sorensen, H. T. Risk of ulcerative colitis and Crohn's disease among offspring of patients with chronic inflammatory bowel disease. *The American journal of gastroenterology* **94**, 3236-3238, doi:10.1111/j.1572-0241.1999.01526.x (1999).

12 Spehlmann, M. E. *et al.* Epidemiology of inflammatory bowel disease in a German twin cohort: results of a nationwide study. *Inflammatory bowel diseases* **14**, 968-976, doi:10.1002/ibd.20380 (2008).

13 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

14 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

15 DePristo, M. A. *et al.* A framework for variation discovery and genotyping

using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498, doi:10.1038/ng.806 (2011).

16 Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **11**, 11 10 11-11 10 33, doi:10.1002/0471250953.bi1110s43 (2013).

17 Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

18 Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-69, doi:10.1126/science.1219240 (2012).

19 Stenson, P. D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome medicine* **1**, 13, doi:10.1186/gm13 (2009).

20 Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Human mutation* **21**, 577-581, doi:10.1002/humu.10212 (2003).

21 Amberger, J., Bocchini, C. & Hamosh, A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM(R)). *Human mutation* **32**, 564-567, doi:10.1002/humu.21466 (2011).

22 Das, J. & Yu, H. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC systems biology* **6**, 92, doi:10.1186/1752-0509-6-92 (2012).

23 Michail, S., Bultron, G. & Depaolo, R. W. Genetic variants associated with Crohn's disease. *The application of clinical genetics* **6**, 25-32, doi:10.2147/TACG.S33966 (2013).

24 O'Neill, L. A., Golenbock, D. & Bowie, A. G. The history of Toll-like receptors - redefining innate immunity. *Nature reviews. Immunology* **13**, 453-460, doi:10.1038/nri3446 (2013).

25 Kellogg, E. H., Leaver-Fay, A. & Baker, D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **79**, 830-838, doi:10.1002/prot.22921 (2011).

26 Palumbo, A. & Anderson, K. Multiple myeloma. *The New England journal of medicine* **364**, 1046-1060, doi:10.1056/NEJMra1011442 (2011).

27 Kuehl, W. M. & Bergsagel, P. L. Multiple myeloma: evolving genetic events and host interactions. *Nature reviews. Cancer* **2**, 175-187, doi:10.1038/nrc746 (2002).

28 Raab, M. S., Podar, K., Breitkreutz, I., Richardson, P. G. & Anderson, K. C. Multiple myeloma. *Lancet* **374**, 324-339, doi:10.1016/S0140-6736(09)60221-X (2009).

29 Lynch, H. T., Sanger, W. G., Pirruccello, S., Quinn-Laquer, B. & Weisenburger, D. D. Familial multiple myeloma: a family study and review of the literature. *Journal of the National Cancer Institute* **93**, 1479-1483 (2001).

30 Vachon, C. M. *et al.* Increased risk of monoclonal gammopathy in first-degree relatives of patients with multiple myeloma or monoclonal gammopathy of undetermined significance. *Blood* **114**, 785-790, doi:10.1182/blood-2008-12-192575 (2009).

31 Eriksson, M. & Hallberg, B. Familial occurrence of hematologic malignancies and other diseases in multiple myeloma: a case-control study. *Cancer causes & control : CCC* **3**, 63-67 (1992).
32 Lynch, H. T. & Thome, S. D. Familial multiple myeloma. *Blood* **114**, 749-750, doi:10.1182/blood-2009-03-207233 (2009).
33 Lynch, H. T. *et al.* Familial myeloma. *The New England journal of medicine* **359**, 152-157, doi:10.1056/NEJMoa0708704 (2008).
34 Rappaport, N. *et al.* MalaCards: A Comprehensive Automatically-Mined Database of Human Diseases. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **47**, 1 24 21-21 24 19, doi:10.1002/0471250953.bi0124s47 (2014).
35 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27-30 (2000).
36 Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic acids research* **42**, D472-477, doi:10.1093/nar/gkt1102 (2014).
37 Milacic, M. *et al.* Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers* **4**, 1180-1211, doi:10.3390/cancers4041180 (2012).
38 Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research* **42**, D1001-1006, doi:10.1093/nar/gkt1229 (2014).
39 Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467-472, doi:10.1038/nature09837 (2011).
40 Kan, Z. *et al.* Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* **466**, 869-873, doi:10.1038/nature09208 (2010).
41 Dees, N. D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome research* **22**, 1589-1598, doi:10.1101/gr.134635.111 (2012).
42 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).
43 Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E455-464, doi:10.1073/pnas.1322563111 (2014).
44 Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185-190, doi:10.1038/nature12975 (2014).
45 Robinson, J. T. *et al.* Integrative genomics viewer. *Nature biotechnology* **29**, 24-26, doi:10.1038/nbt.1754 (2011).
46 Carmi, S. *et al.* Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. *Nature communications* **5**, 4835, doi:10.1038/ncomms5835 (2014).

APPENDIX A

Supplementary Information for Chapter 1

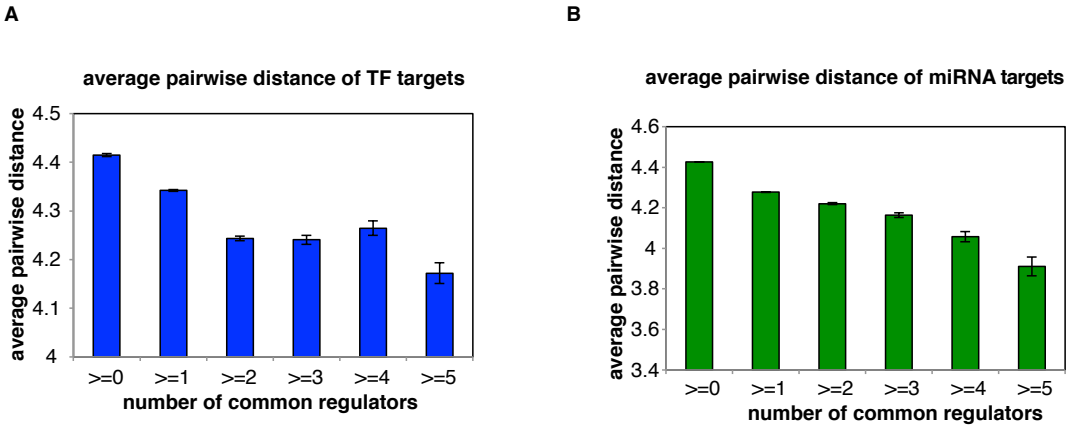


Figure A.1 Synergistic effect of TF and miRNA regulation on the protein interactions of their targets. Average pairwise distance of protein products of gene targets regulated by multiple TFs **(A)** or miRNAs **(B)** in the protein-protein interaction network.

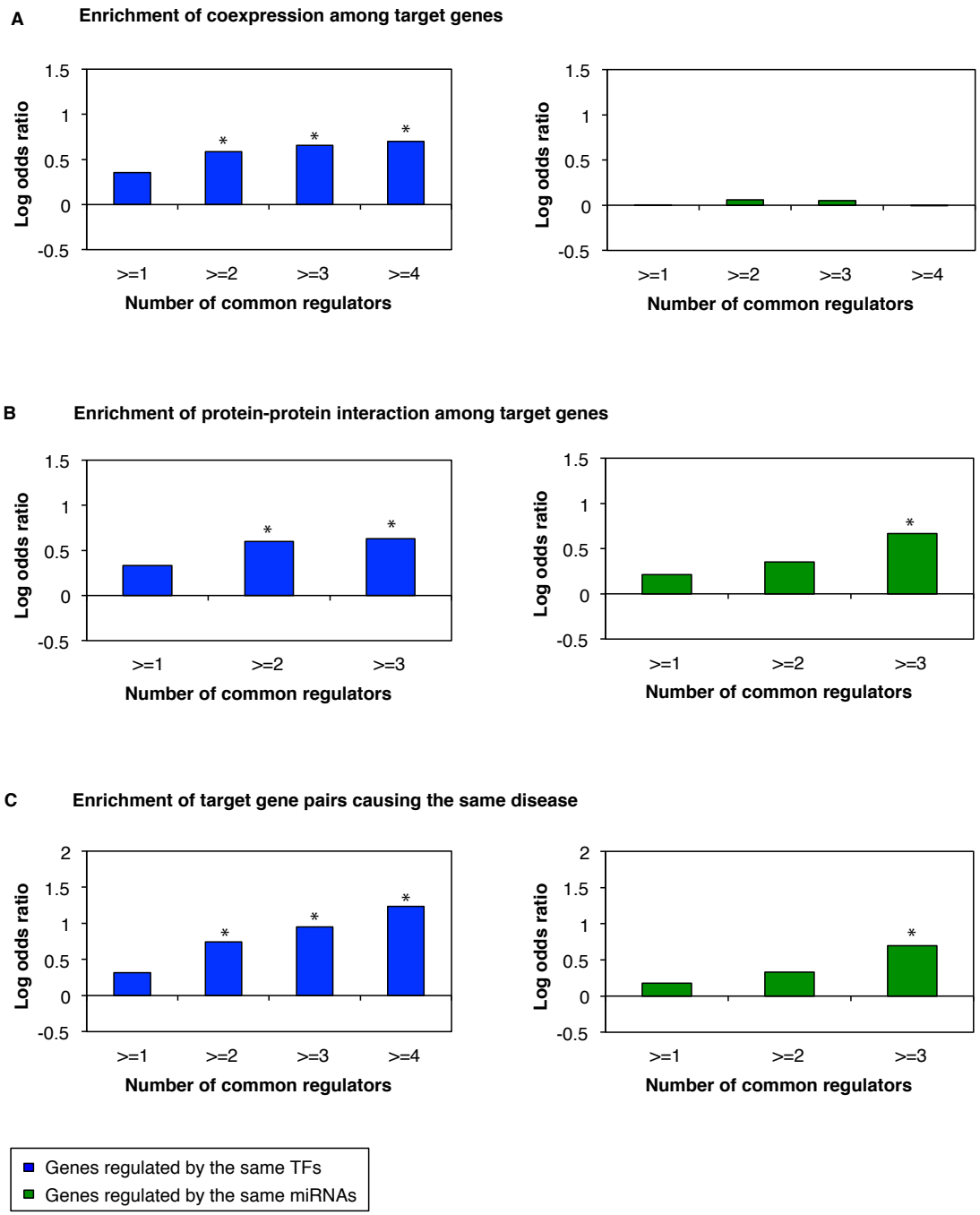


Figure A.2 Analysis of synergistic effects of TFs and miRNAs on target genes using 100 randomized networks as control. **(A)** LOD values for enrichment of co-expression relationships between gene pairs jointly regulated by multiple TFs (left) or miRNAs (right). **(B)** LOD values for enrichment of protein-protein interactions between gene pairs jointly regulated by multiple TFs (left) or miRNAs (right). **(C)** LOD values for the likelihood of gene pairs jointly regulated by multiple TFs (left) or miRNAs (right) to be associated with the same disease. *$P < 0.05$. *P*-values are calculated by the *Z*-test.

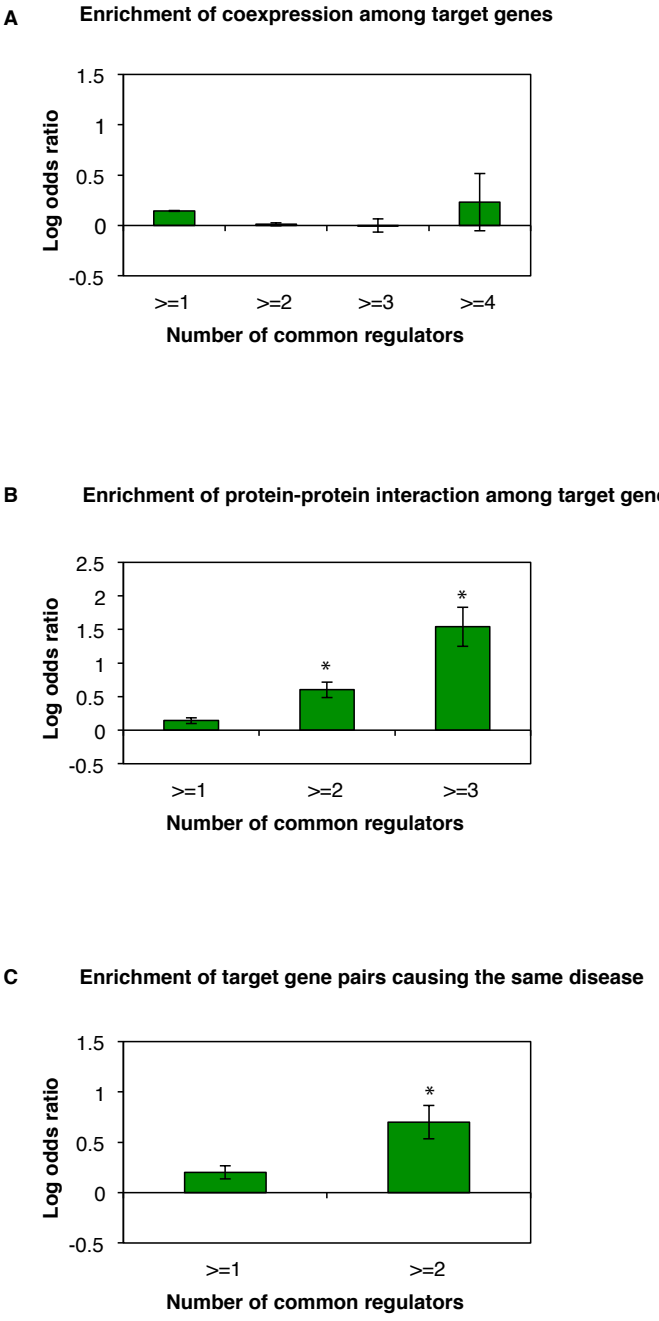


Figure A.3 Analysis of the synergistic effects of miRNAs on target genes using experimentally verified miRNA targets. **(A)** LOD values for enrichment of co-expression relationships between gene pairs jointly regulated by multiple miRNAs. **(B)** LOD values for enrichment of protein-protein interactions between gene pairs jointly regulated by miRNAs. **(C)** LOD values for the likelihood of gene pairs jointly regulated by multiple miRNAs to be associated with the same disease. *$P < 0.05$. *P*-values are calculated by the *Z*-test.

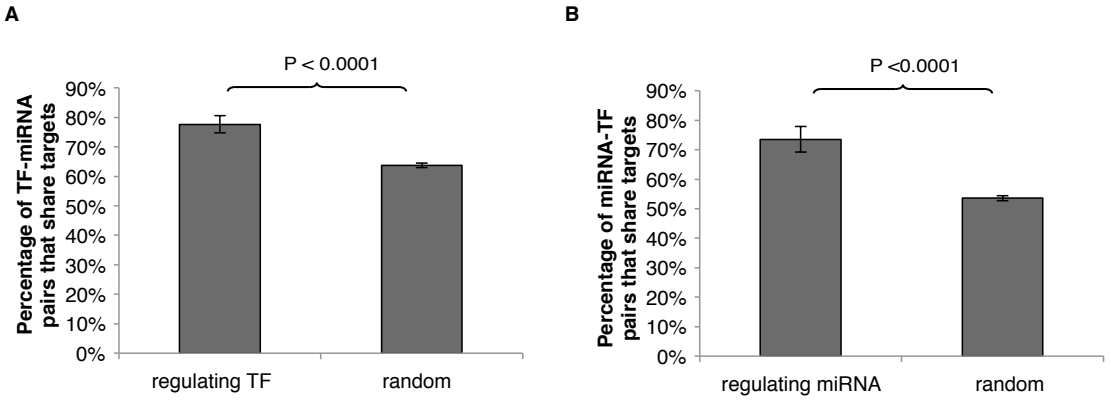


Figure A.4 Co-regulation of gene targets by inter-regulating TFs and miRNAs. **(A)** Percentage of miRNAs that share targets with TFs that regulate them. **(B)** Percentage of TFs that share targets with miRNAs that regulate them. Error bars indicate ±SE.

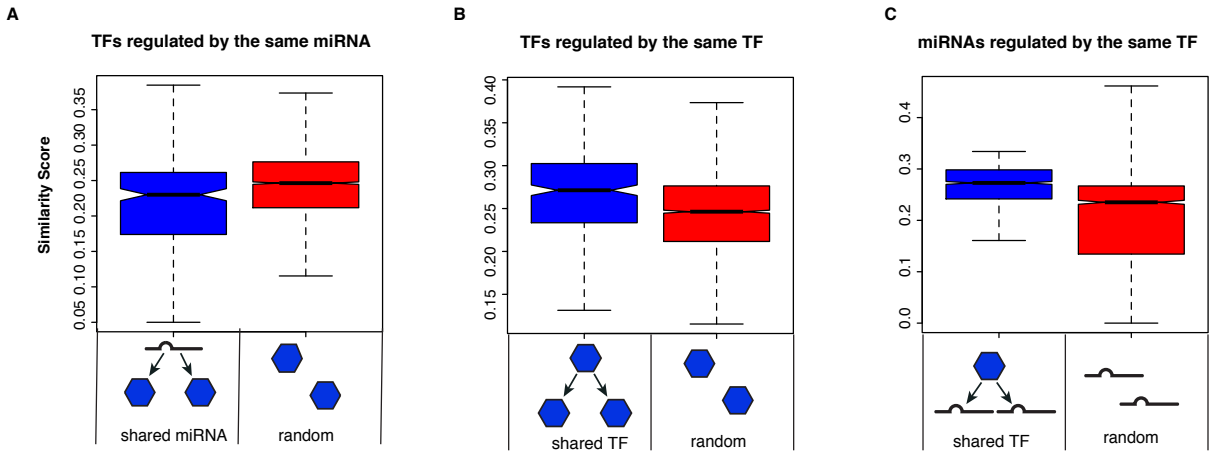


Figure A.5 Inter-regulation of TFs and miRNAs. Distribution of the similarity scores of **(A)** TF pairs regulated by the same miRNA, **(B)** TF pairs regulated by the same TF, and **(C)** miRNA pairs regulated by the same TF.

APPENDIX B

Supplementary Information for Chapter 2

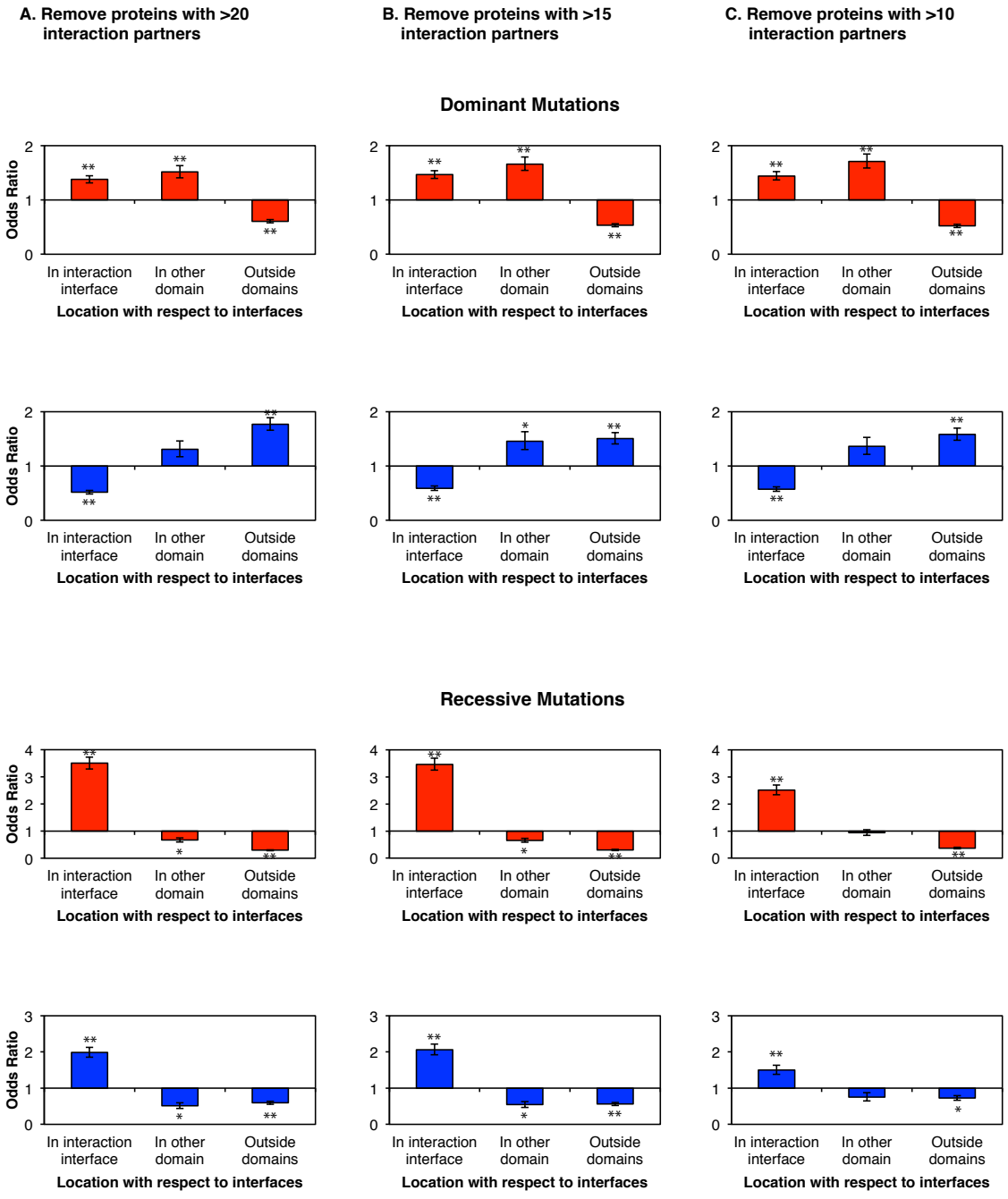


Figure B.1 Distribution of recessive and dominant disease mutations with respect to interaction interfaces after removal of protein hubs. Error bars represent 95% confidence intervals of odds ratios. $^{**}P < 10^{-20}$, $^{*}P < 10^{-10}$. *P*-values are calculated using *Z*-tests for log odds ratio. Red: in-frame mutations. Blue: truncating mutations.

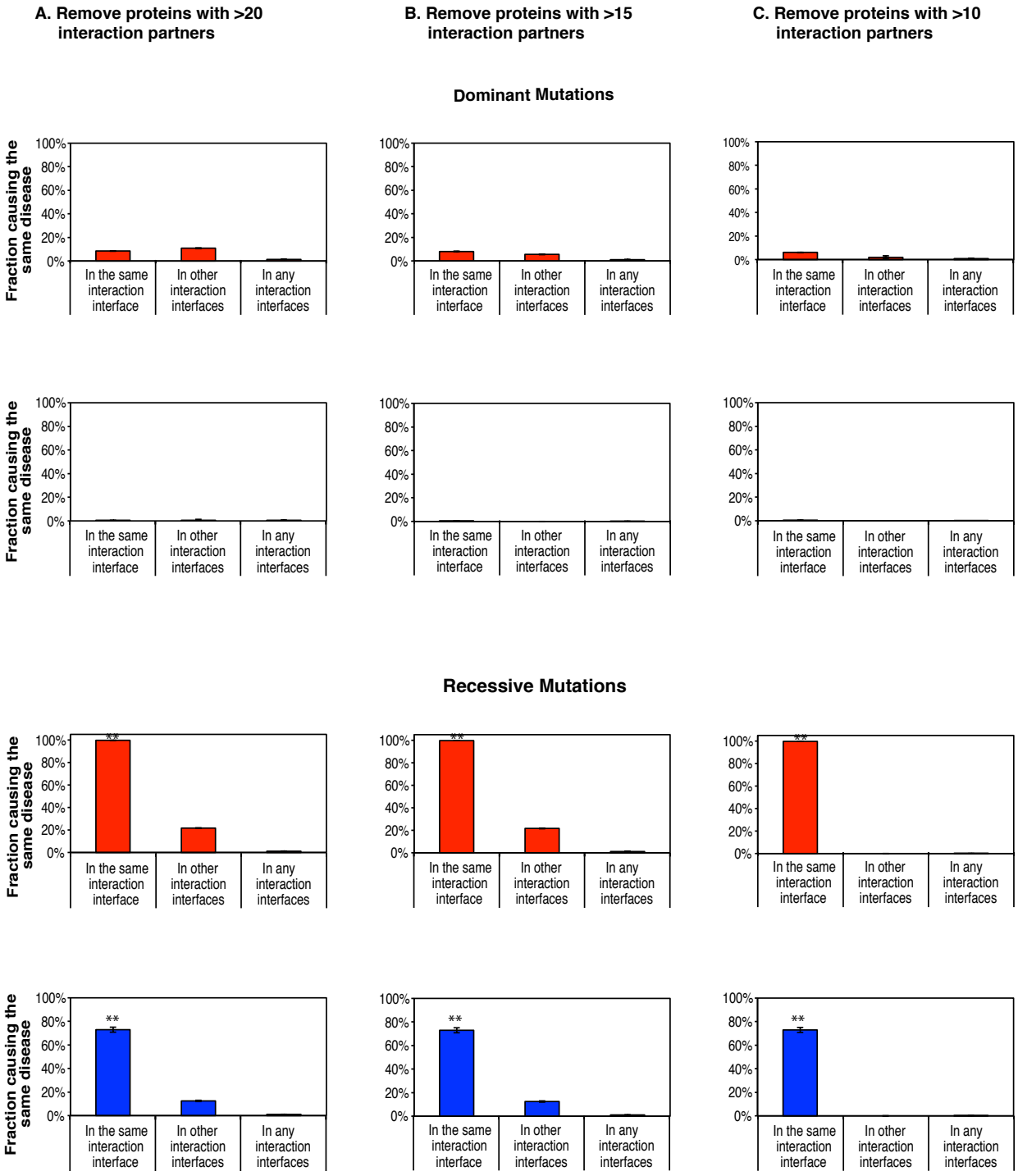


Figure B.2 Analysis of locus heterogeneity among dominant and recessive disease mutations after removal of protein hubs. Error bars represent ±SE. $^{**}P < 10^{-20}$. *P*-values are calculated using cumulative binomial tests. Red: in-frame mutations. Blue: truncating mutations.

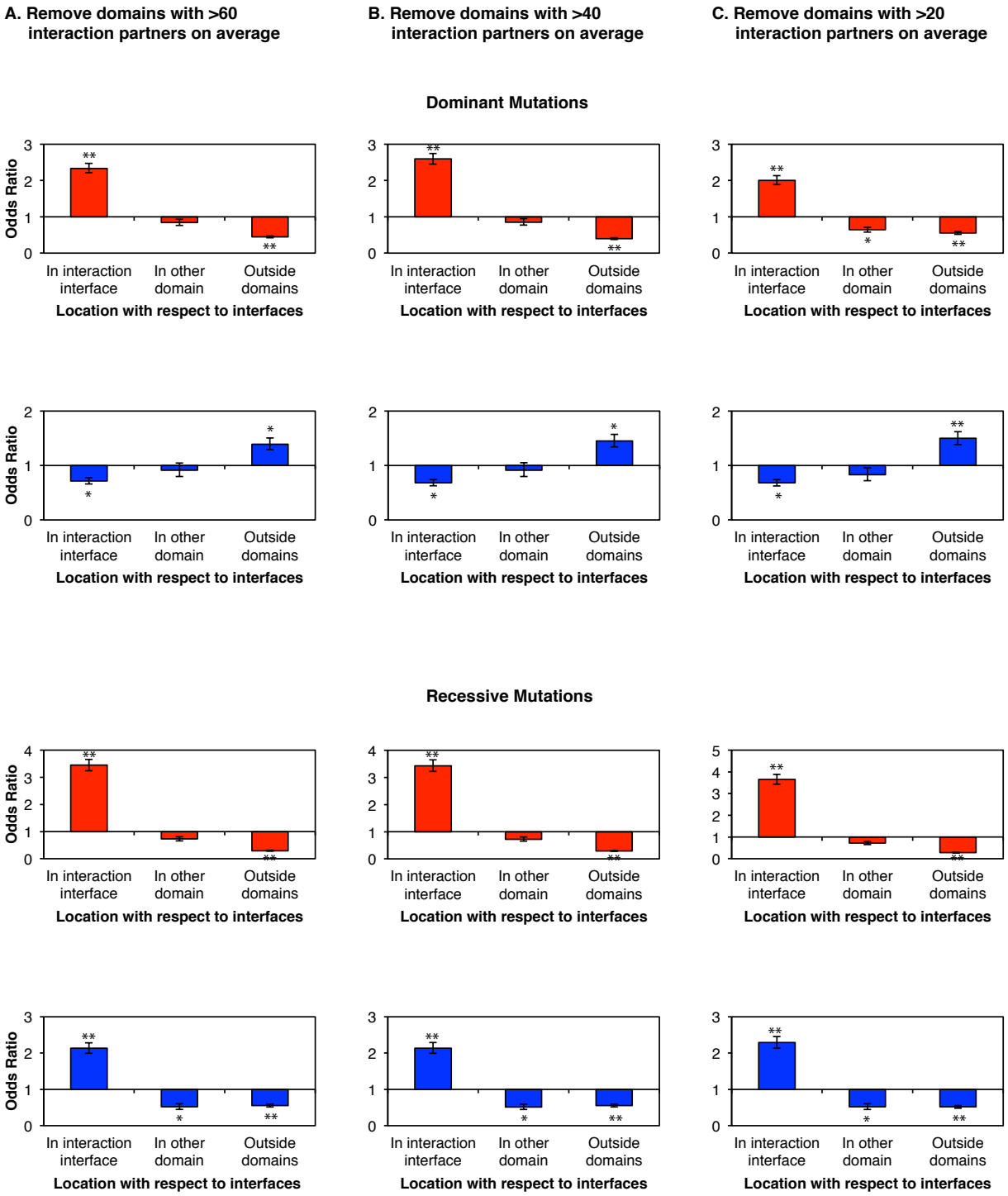


Figure B.3 Distribution of recessive and dominant disease mutations with respect to interaction interfaces after removal of domain hubs. Error bars represent 95% confidence intervals of odds ratios. $^{**}P < 10^{-20}$, $^{*}P < 10^{-10}$. *P*-values are calculated using *Z*-tests for log odds ratio. Red: in-frame mutations. Blue: truncating mutations. Note that the enrichment of dominant mutations in other domains decreased after the removal of domain hubs, suggesting that this enrichment might be due to over-represented domains in the 3D protein interactome network.

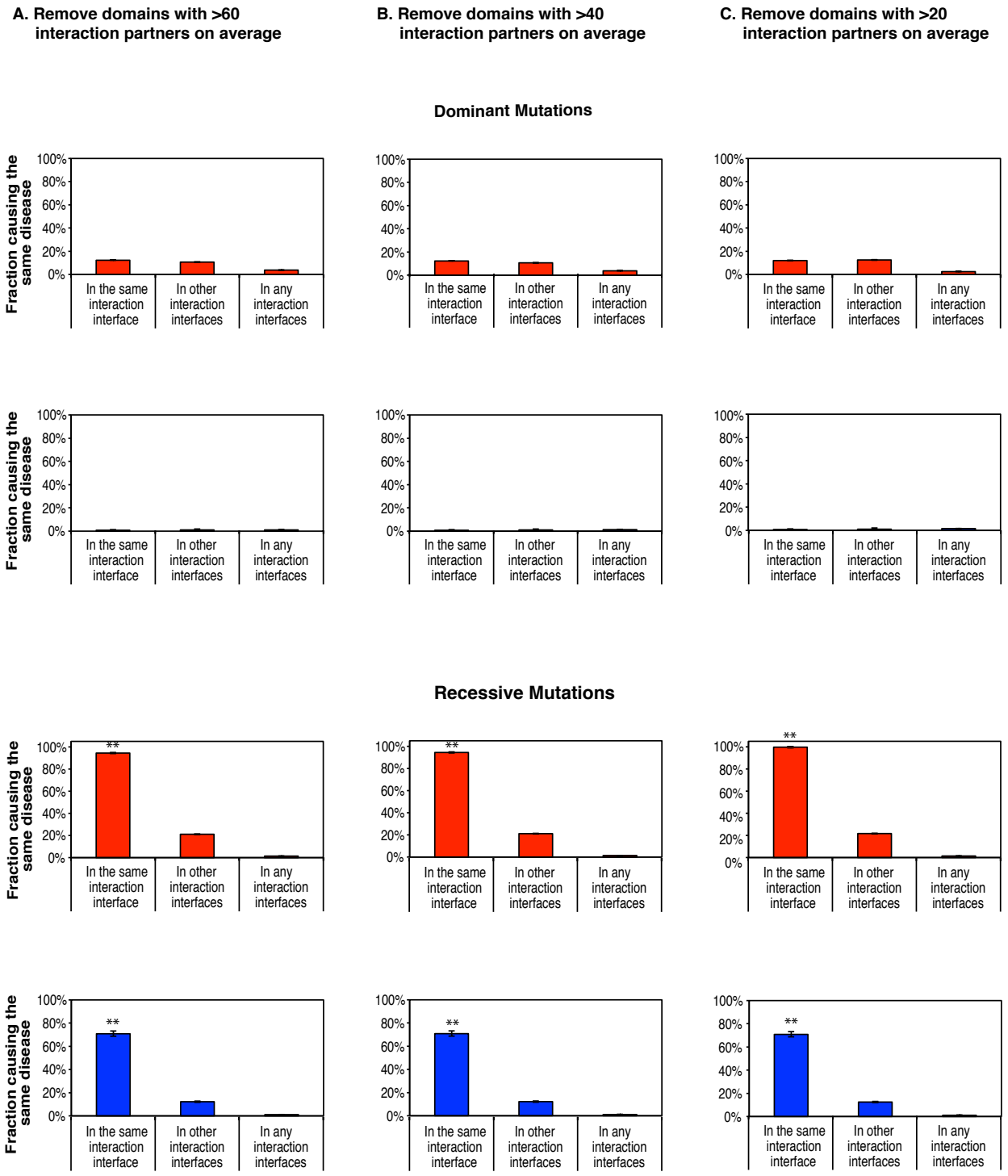
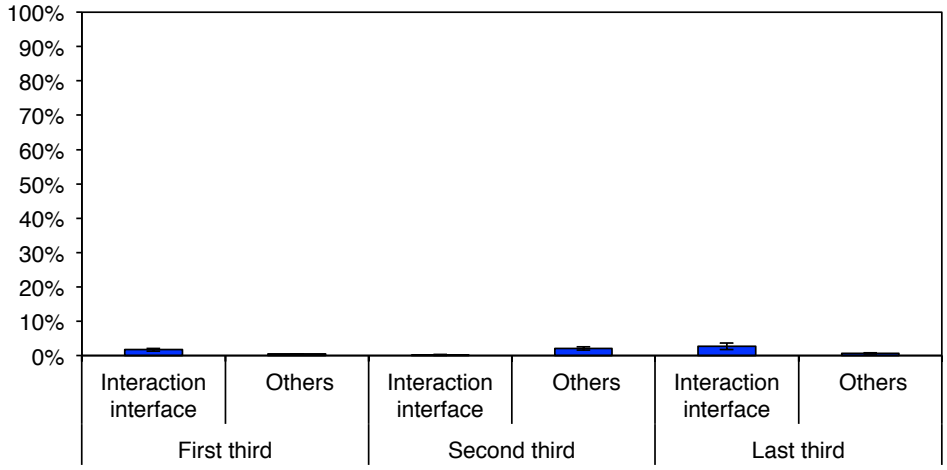


Figure B.4 Analysis of locus heterogeneity among dominant and recessive disease mutations after removal of domain hubs. Error bars represent ±SE. $^{**}P < 10^{-20}$. *P*-values are calculated using cumulative binomial tests. Red: in-frame mutations. Blue: truncating mutations.

A



B

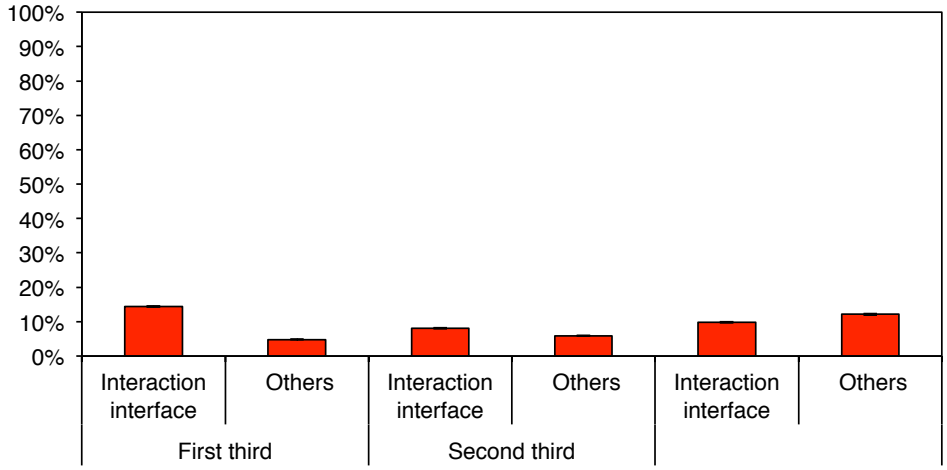


Figure B.5 Analysis of the effect of mutation location on locus heterogeneity of dominant disease mutations. **(A)** Percentage of dominant truncating mutations located on different parts of the protein that cause the same disease with mutations on its interaction partner. **(B)** Percentage of dominant in-frame mutations located on different parts of the protein that cause the same disease with mutations on its interaction partner.

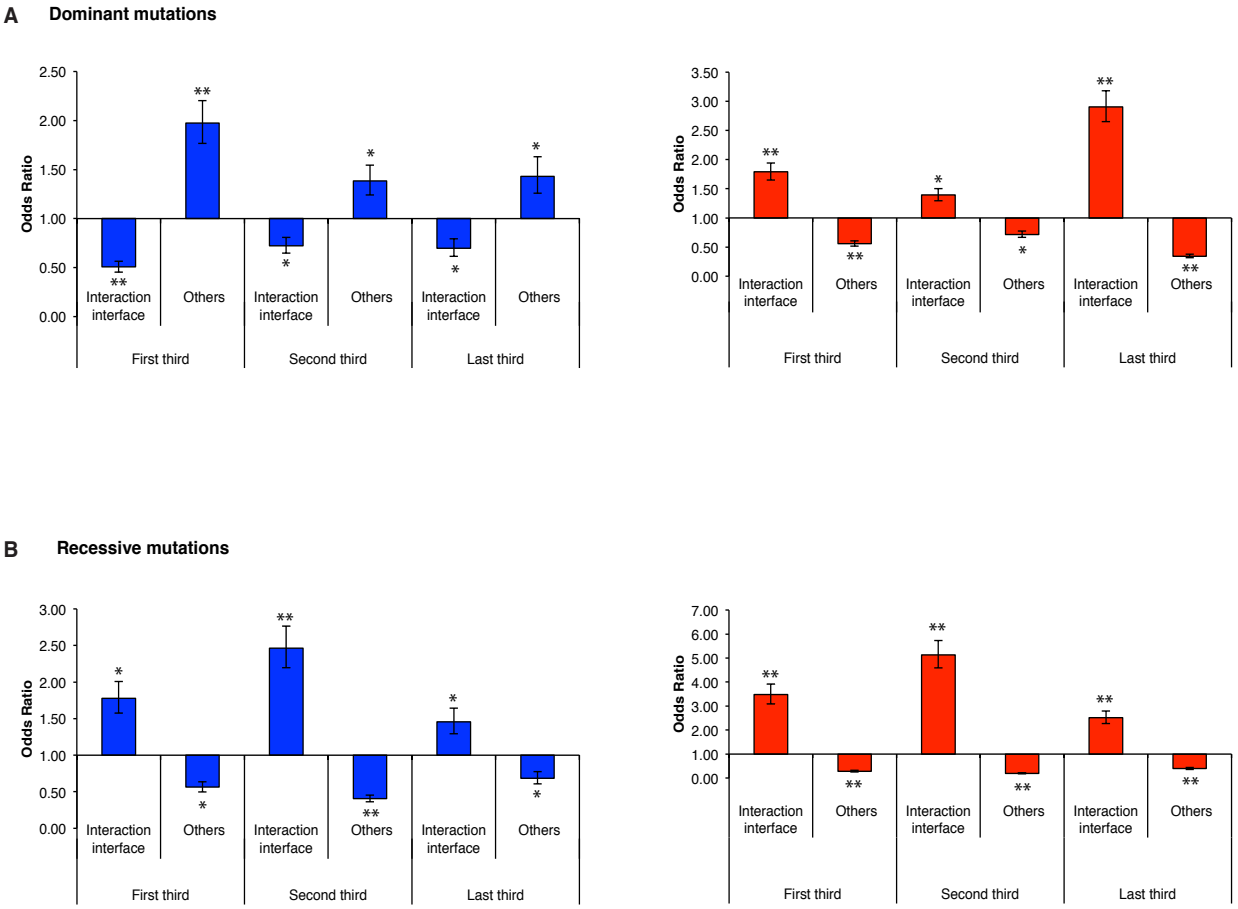
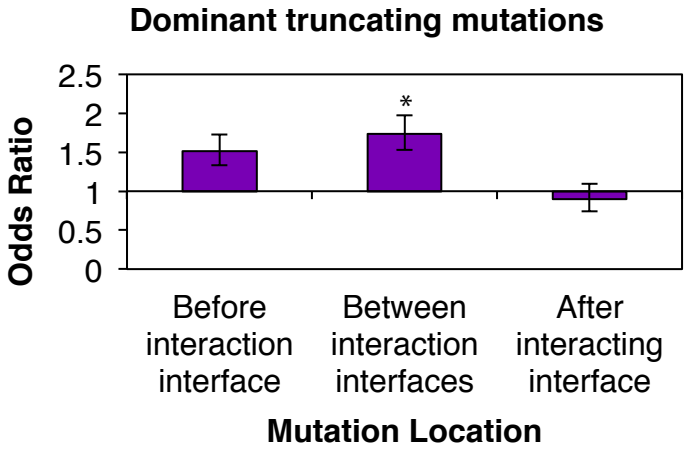


Figure B.6 Effect of mutation location relative to the N-terminus on the mutation distribution patterns. **(A)** Odds ratios of the distributions of dominant truncating (left) and in-frame (right) mutations on different locations of proteins. **(B)** Odds ratios of the distributions of recessive truncating (left) and in-frame (right) mutations on different locations of proteins. Error bars represent 95% confidence intervals of odds ratios. $^{**}P < 10^{-20}$, $^{*}P < 10^{-5}$. *P*-values are calculated using *Z*-tests for log odds ratio.

A



B

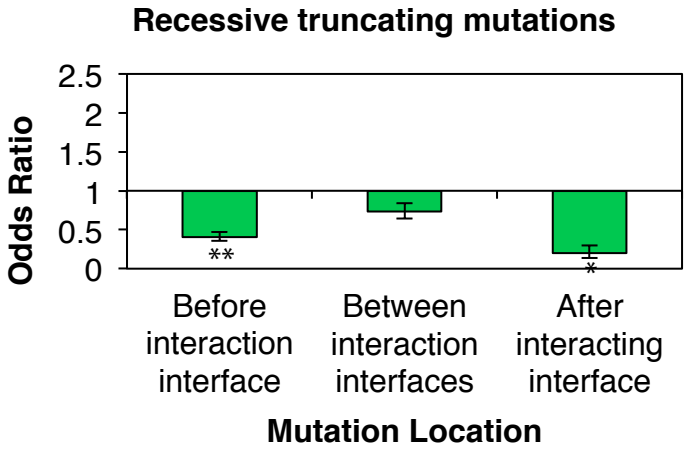


Figure B.7 Distribution of recessive and dominant truncating mutations outside of interaction interfaces. **(A)** Odds ratios of the distribution of dominant truncating mutations outside of interaction interfaces. **(B)** Odds ratios of the distribution of recessive truncating mutations outside of interaction interfaces. Error bars represent 95% confidence intervals of odds ratios. $^{**}P < 10^{-20}$, $^{*}P < 10^{-10}$. *P*-values are calculated using *Z*-tests for log odds ratio.

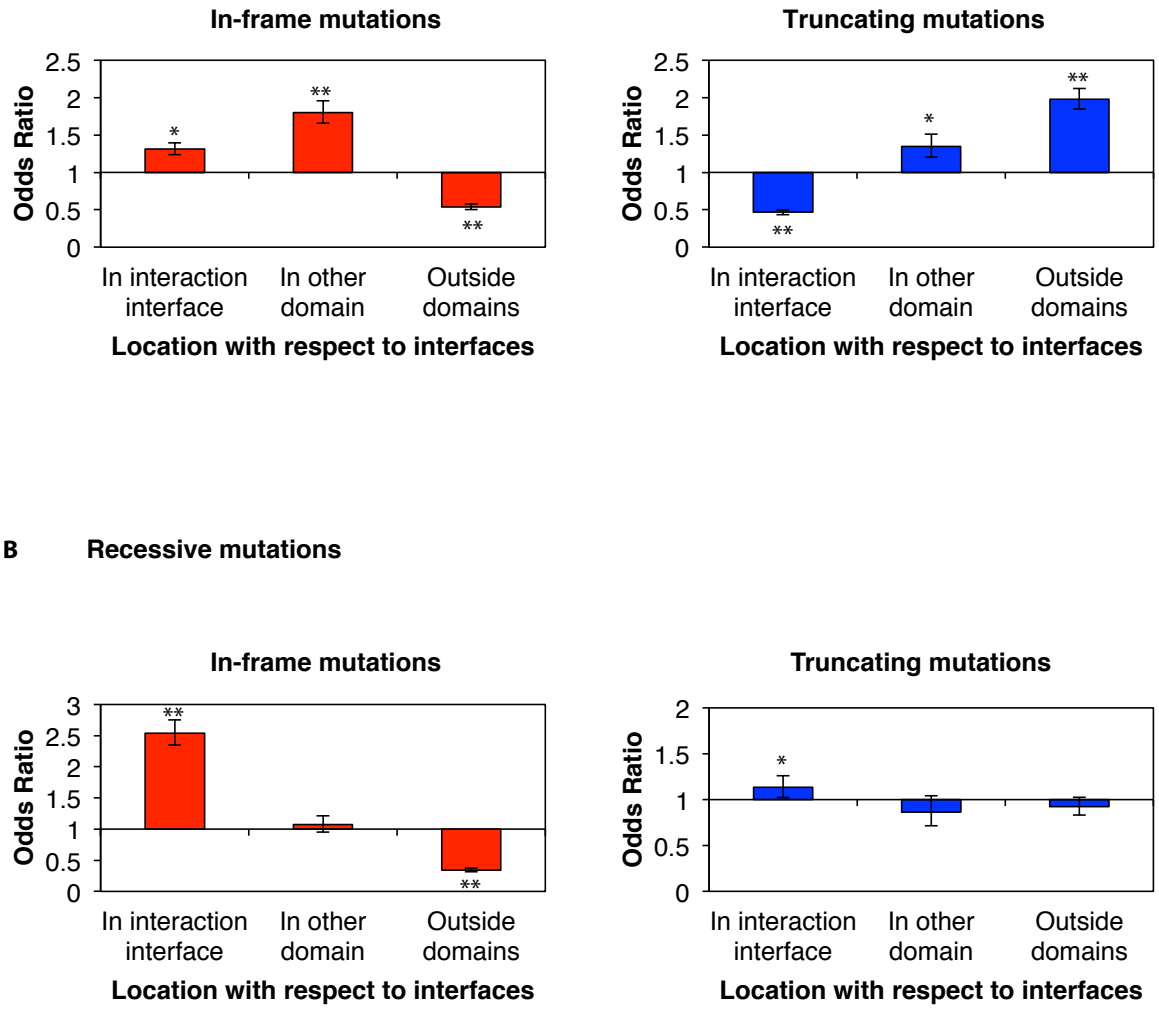


Figure B.8 Distribution of recessive and dominant disease mutations from HGMD with respect to interaction interfaces. **(A)** Odds ratios of the distributions of dominant in-frame (left) and truncating (right) mutations on different locations of proteins. **(B)** Odds ratios of the distribution of recessive in-frame (left) and truncating (right) mutations on different locations of proteins. Error bars represent 95% confidence intervals of odds ratios. $^{**}P < 10^{-20}$, $^{*}P < 0.05$. *P*-values are calculated using *Z*-tests for log odds ratio.

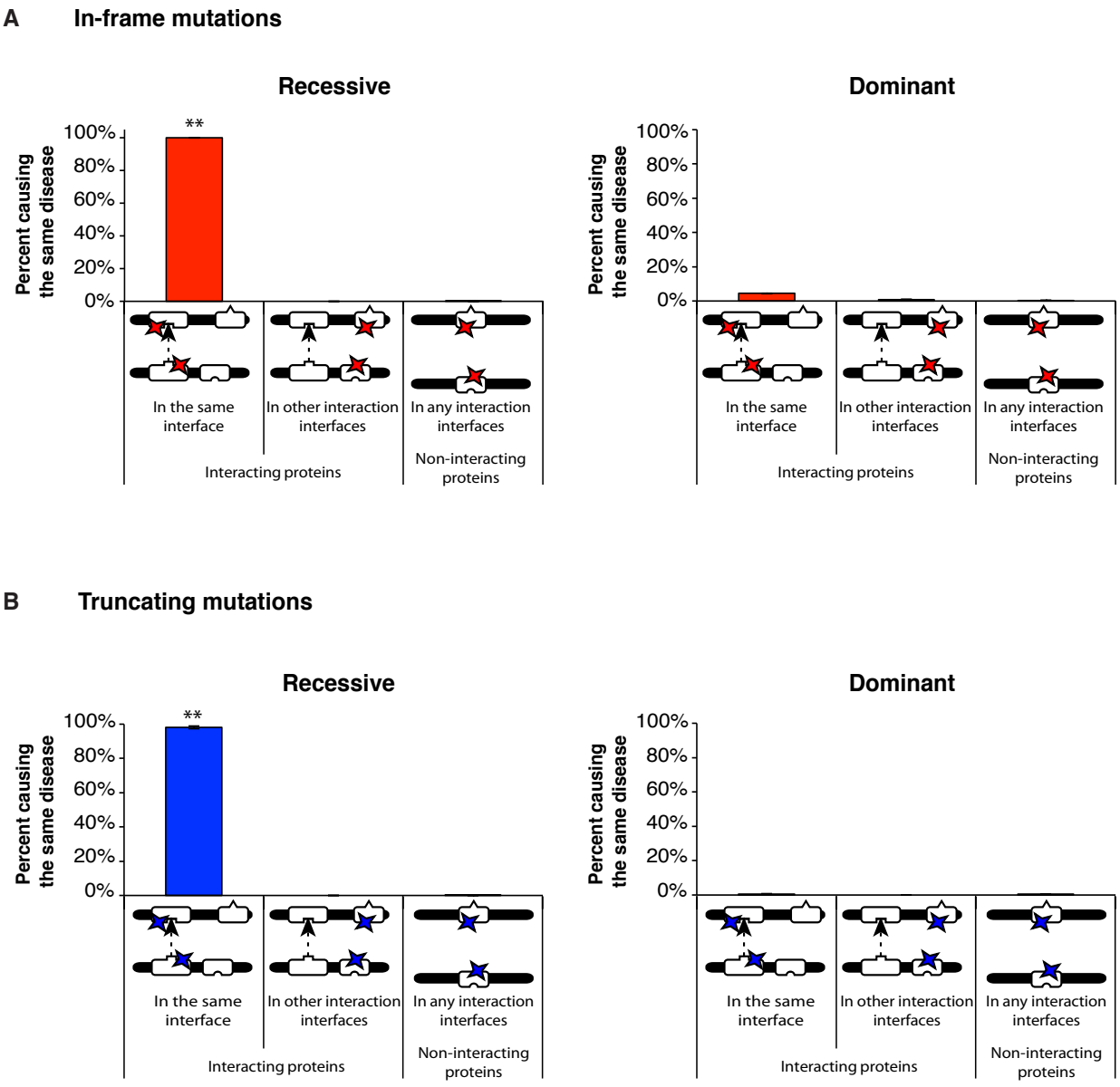


Figure B.9 Analysis of locus heterogeneity among dominant and recessive HGMD disease mutations. **(A)** Percentage of recessive (left) or dominant (right) in-frame mutation pairs on two different proteins causing the same disease. **(B)** Percentage of recessive (left) or dominant (right) truncating mutation pairs on two different proteins causing the same disease. Error bars represent ±SE. $^{**}P < 10^{-20}$. *P*-values are calculated using cumulative binomial tests.

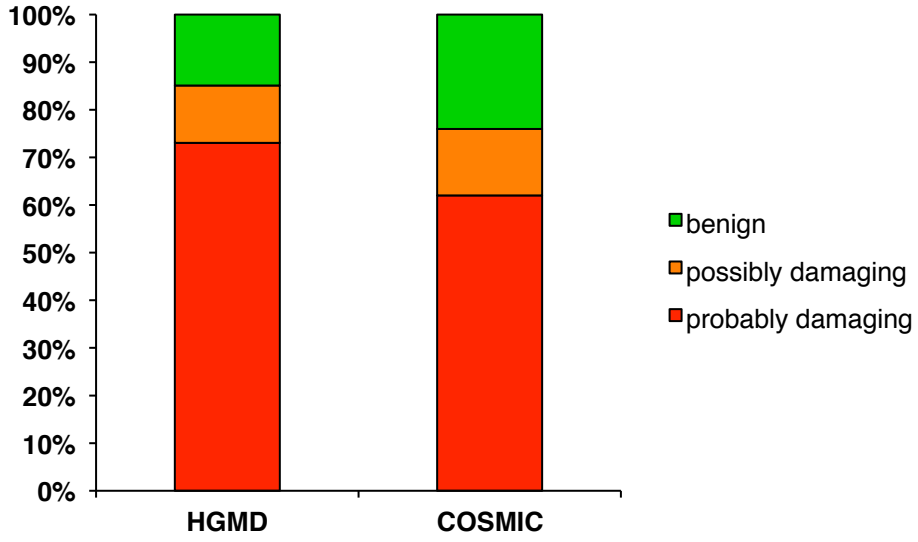


Figure B.10 Polyphen-2 predictions on the missense mutations used for the analyses (HumVar model).

Table B.1 Network statistics of the 3D protein interactome network

Average Degree	**Clustering Coefficient**	**Characteristic Path Length**	**Diameter**
2.73	0.216	8.7	27

Table B.2 Sample sizes used in the calculations

Figure 2.2 Distribution of dominant and recessive mutations

Location of mutations	Dominant		Recessive	
	In-frame	Truncating	In-frame	Truncating
In interaction interfaces	4421	1398	3620	2119
In other domains	764	355	410	188
Outside domains	2339	2013	1122	1243

Figure 2.3A Locus heterogeneity of recessive in-frame mutations

Location of mutation pairs on the interacting proteins		Same disease	Different diseases
Interacting genes	Same interface	2145	296
	Different interfaces	3158	11882
Non-interacting genes	Any interfaces	69832	5257750

Figure 2.3A Locus heterogeneity of dominant in-frame mutations

Location of mutation pairs on the interacting proteins		Same disease	Different diseases
Interacting genes	Same interface	7138	63539
	Different interfaces	2490	21066
Non-interacting genes	Any interfaces	154193	8273199

Figure 2.3B Locus heterogenity of recessive truncating mutations

Location of mutation pairs on the interacting proteins		Same disease	Different diseases
Interacting genes	Same interface	307	197
	Different interfaces	484	3569
Non-interacting genes	Any interfaces	20010	1758000

Figure 2.3B Locus heterogeneity of dominant truncating mutations

Location of mutation pairs on the interacting proteins		Same disease	Different diseases
Interacting genes	Same interface	6	1282
	Different interfaces	1	168
Non-interacting genes	Any interfaces	4076	864499

Figure 2.4A Locus heterogeneity of haploinsufficient in-frame mutations

Location of mutation pairs on the interacting proteins		Same disease	Different diseases
Interacting genes	Same interface	1909	7275
	Different interfaces	772	5711
Non-interacting genes	Any interfaces	25823	1233426

Figure 2.4A Locus heterogeneity of non-haploinsufficient in-frame mutations

Location of mutation pairs on the interacting proteins		Same disease	Different diseases
Interacting genes	Same interface	2360	22805
	Different interfaces	938	8014
Non-interacting genes	Any interfaces	42080	2902891

Figure 2.4B Distribution of haploinsufficient (HI) vs non-haploinsufficient (non-HI) truncating mutations

Location of mutations	HI	non-HI
In interaction interfaces	516	882
In other domains	113	242
Outside domains	401	1611

Fig 2.5A Locus heterogeneity of truncating recessive mutations in thirds

Location of mutation pairs on the interacting proteins		Same disease	Different diseases
First third	In interaction interface	170	150
	Others	277	1828
Second third	In interaction interface	242	114
	Others	160	1127
Last third	In interaction interface	208	182
	Others	115	367

Fig 2.5B Locus heterogeneity of in-frame recessive mutations in thirds

Location of mutation pairs on the interacting proteins		Same disease	Different diseases
First third	In interaction interface	936	103
	Others	1429	2448
Second third	In interaction interface	2193	240
	Others	540	1481
Last third	In interaction interface	1525	309
	Others	432	614

Fig 2.6A Enrichment of truncating mutations in between interacting interfaces

	Dominant	Recessive
Between interacting interfaces	302	265

APPENDIX C

Supplementary Information for Chapter 3

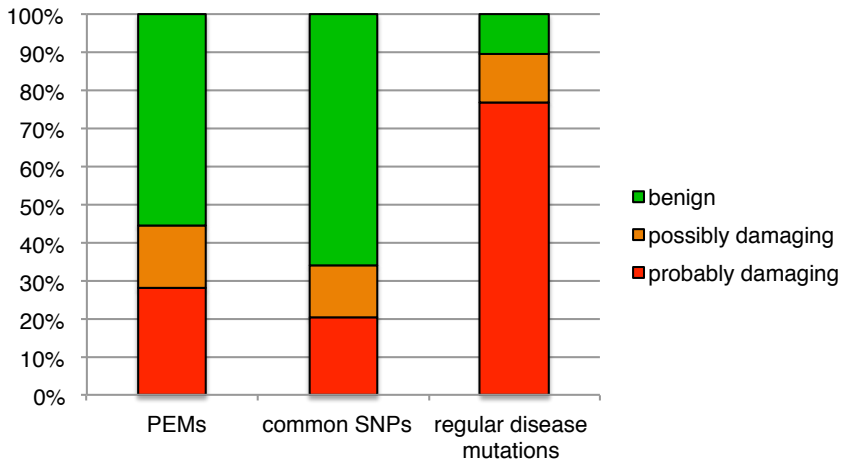


Figure C.1 Polyphen2 predictions on PEMs, common SNPs and regular disease mutations.

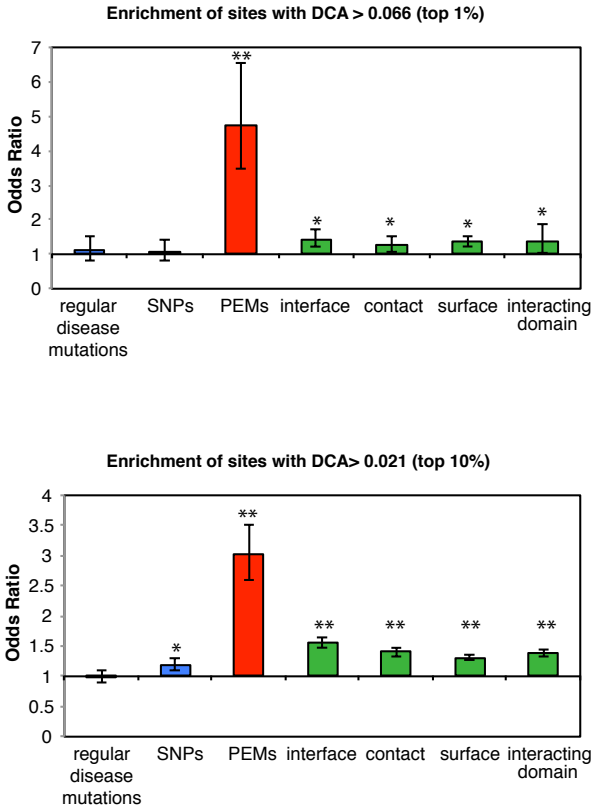


Figure 2 Odds ratios of the enrichment in inter-molecular coevolution between different categories of amino acid. **(A)** Enrichment in sites with DCA score > 0.066 (top 1%). **(B)** Enrichment in sites with DCA score > 0.021 (top 10%). $^{**}P < 10^{-20}$, $^{*}P < 0.05$. *P*-values are calculated using the Z-test.

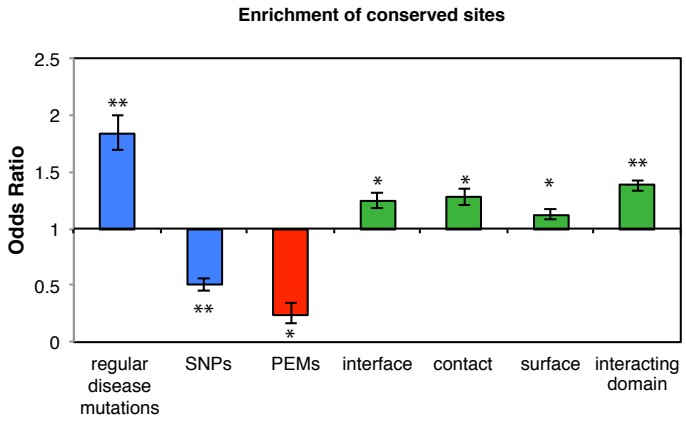


Figure C.3 Odds ratios of the enrichment of conserved sites in different categories of amino acid sites. $^{**}P < 10^{-20}$, $^{*}P < 0.05$. *P*-values are calculated using the Z-test.