Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders

Graphical Abstract

Highlights

- Most missense disease mutations appear not to impair protein folding or stability
- Interaction profiling helps distinguish disease mutations from non-disease variants
- Distinct interaction perturbations underlie distinct disease phenotypes
- Integrative interaction networks enhance genotype-to-phenotype understanding

Authors
Nidhi Sahni, Song Yi, ..., Susan Lindquist, Marc Vidal

Correspondence
lindquist@wi.mit.edu (S.L.), marc_vidal@dfci.harvard.edu (M.V.)

In Brief
A large-scale characterization of disease mutations reveals surprisingly widespread, yet specific perturbations in macromolecular interactions. Different mutations in the same gene lead to different interaction profiles, often resulting in distinct disease phenotypes.

Sahni et al., 2015, Cell 161, 647–660
April 23, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2015.04.013
Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders


SUMMARY

How disease-associated mutations impair protein activities in the context of biological networks remains mostly undetermined. Although a few renowned alleles are well characterized, functional information is missing for over 100,000 disease-associated variants. Here we functionally profile several thousand missense mutations across a spectrum of Mendelian disorders using various interaction assays. The majority of disease-associated alleles exhibit wild-type chaperone binding profiles, suggesting they preserve protein folding or stability. While common variants from healthy individuals rarely affect interactions, two-thirds of disease-associated alleles perturb protein-protein interactions, with half corresponding to “edgetic” alleles affecting only a subset of interactions while leaving most other interactions unperturbed. With transcription factors, many alleles that leave protein-protein interactions intact affect DNA binding. Different mutations in the same gene leading to different interaction profiles often result in distinct disease phenotypes. Thus disease-associated alleles that perturb distinct protein activities rather than grossly affecting folding and stability are relatively widespread.

INTRODUCTION

Over a hundred thousand genetic variants have been identified across a large number of Mendelian disorders (Amberger et al., 2011), complex traits (Hindorff et al., 2009), and cancer types...
However, many fundamental questions regarding genotype-phenotype relationships remain unresolved (Vidal et al., 2011). One critical challenge is to distinguish causal disease mutations from non-pathogenic polymorphisms. Even when causal mutations are identified, the functional consequence of such mutations is often elusive (Sahni et al., 2013). Genotypic information alone rarely elucidates the mechanistic insights pertaining to disease pathogenesis. Although genotype-phenotype relationships can be modeled under the assumption that most disease-associated mutations lead to complete loss of protein function, e.g., through radical changes such as protein misfolding and instability (Subramanian and Kumar, 2006) (Figure 1A), the reality is often more complex, as in the case of mutations affecting the same gene but giving rise to clinically distinguishable diseases (Zhong et al., 2009). In addition, since genes and gene products do not function in isolation but interact with each other in the context of interactome networks (Vidal et al., 2011), it is likely that many diseases result from perturbations of such complex networks (Goh et al., 2007).

Missense mutations are among the most common sequence alterations in Mendelian disorders, accounting for more than half of all reported mutations in the Human Gene Mutation Database (HGMD) (Stenson et al., 2014). In principle, missense mutations may have no functional consequences, disrupt the three-dimensional structure of the corresponding protein, or exert specific effects on particular molecular or biochemical interactions (Figure 1A), such as protein-protein interactions (PPIs), protein-DNA interactions (PDIs), or enzyme-substrate interactions, while leaving all other functional properties unperturbed. We previously reported that a considerable portion of Mendelian disease mutations could indeed be predicted computationally to cause interaction-specific, or “edgetic,” perturbations (Zhong et al., 2009). However, only a small number of genes and associated mutations were experimentally tested in that study, and the extent to which disease mutations globally lead to interaction perturbations remains to be determined.

Here we describe a multi-pronged approach to systematically decipher molecular interaction perturbations associated with missense mutations. Since chaperones and associated quality control factors (QCFs) can salvage unstable proteins by assisting with folding, and an increase in protein-chaperone interactions (PCIs) has been observed for a number of disease mutants (Whitesell and Lindquist, 2005), our systematic approach begins with characterizing PCIs for large numbers of disease-associated alleles, followed by systematic measurements of PPI and PDIs.
PDI profile changes caused by mutations, a strategy referred to as “edgotyping” (Figure 1B).

We provide evidence for widespread interaction perturbations across a broad spectrum of human Mendelian disorders. Our results suggest that interaction profiling helps distinguish disease-causing mutations from common variants. Furthermore, the integration of different types of molecular interactions expands our ability to understand complex genotype-phenotype relationships.

RESULTS

Human Mutation ORFeome Version 1.1
To globally characterize disease-associated alleles, we selected mutations associated with a wide range of disorders, including cancer susceptibility and heart, respiratory, and neurological diseases. We retrieved from HGMD (Stenson et al., 2014) a list of ~16,400 mutations affecting over 1,200 genes for which we have a wild-type (WT) open-reading frame (ORF) clone in our human “ORFeome” collection (Yang et al., 2011) and selected up to four mutations per gene (Figure 1C; Tables S1A and S1B; Extended Experimental Procedures). Using properties related to RNA abundance, GO annotation, and protein domains (Extended Experimental Procedures), we verified there is no significant bias between our selected genes and the rest of the human genome or all genes represented in HGMD (Figures S1B–S1G).

Altogether, we cloned and sequence-verified 2,890 human mutant ORFs (hmORFs), each harboring a single nucleotide change that results in an amino acid change relative to the corresponding WT ORF of 1,140 genes. To our knowledge, this human mutation ORFeome version 1.1 resource (hmORFeome1.1; Figure S1A) is the most extensive human mutation collection reported to date.

Disease Mutations and Protein Folding and Stability
Using enhanced binding to a chaperone as an indicator of protein instability or misfolding, we examined how disease mutations impact protein folding and disposition. We determined the extent to which hmORF-encoded proteins and their WT counterparts interact with QCFs using a quantitative high-throughput LUMIER assay (Taipale et al., 2012; Taipale et al., 2014) (Figure 1C and Table S2A). We selected the following QCFs based on their broad specificity (Taipale et al., 2014): (1) the cytoplasmic chaperones HSP90 and HSC70, (2) their co-chaperones BAG2 and CHIP/STUB1, (3) the proteasomal regulatory subunit PSMD2 (formerly known as RPN1), and (4) the ER chaperones GRP78/BIP and GRP94 (Extended Experimental Procedures). We did not survey mitochondrial chaperones since only ~7% of disease-associated gene products are predicted to localize solely in mitochondria (Huntley et al., 2015).

Increased interaction between a QCF and mutant or WT protein, as measured by the LUMIER assay, indicates a mutation-induced perturbation in conformational stability, often associated with compromised or complete loss of function (Taipale et al., 2012). The interaction profiles of most mutant proteins correlated with their WT counterparts. However, compared to a background control set, we observed a significant enrichment of mutant alleles showing increased interaction with QCFs (Figures 2A–2H and S2A) but little or no enrichment for decreased interaction (Figures 2A and S2B; Extended Experimental Procedures). The interaction profiles of mutant proteins with the different cytoplasmic QCFs were highly correlated, distinct from those with ER factors (Figure 2I). These results highlight the coordination and specificity of cellular quality control pathways. Altogether ~28% of the tested alleles exhibited increased binding to at least one of the seven QCFs tested. Although this fraction is likely a conservatively low estimate due to limited assay sensitivity, the strong correlation between chaperone interaction profiles (Figure 2I) suggests that the estimate would not increase substantially by assaying more chaperones. We validated several mutant-specific interactions with endogenous chaperones by co-immunoprecipitation followed by western blot, corroborating the results obtained with the LUMIER assay (Figure 2J).

We next estimated protein abundance using semiquantitative ELISA, which provides a proxy for steady-state protein stability. Although the expression levels of mutant alleles correlated with their WT counterparts (Figure S2C), mutant proteins exhibiting enhanced interactions with cytoplasmic, but not ER, chaperones were detected at lower steady-state levels than their WT counterparts (p < 1.0 × 10⁻⁴; Figure 3A). This is possibly a result of retention in the ER of mutant proteins that would normally be secreted and therefore not be detected by an assay that captures intracellular proteins. Interestingly, recessive alleles exhibited lower protein abundance levels and increased binding with QCFs compared with proteins encoded by dominant alleles (Figures S2D and S2E). This is consistent with the hypothesis that recessive mutations are more likely to result in loss-of-function phenotypes than dominant mutations (Lesage and Brice, 2009).

To gain insight into the structural properties of mutant proteins that exhibit increased binding to QCFs, we assessed the impact of different disease mutations on predicted protein structures. The disease alleles associated with increased binding to QCFs corresponded significantly more often to mutations of residues buried in the core of the protein (Figure 3B and Table S1C), and less often to mutations in intrinsically disordered regions (Figure 3C) when compared to mutant proteins with no change in binding. Next, we estimated the relative “deleteriousness” associated with distinct genetic mutations using PolyPhen-2 algorithm (Adzhubei et al., 2010). Deleterious mutations predicted by PolyPhen were significantly enriched in alleles that exhibited increased binding to QCFs (Figure S2F).

Previous studies suggested that increased chaperone binding reflects a change in protein stability (Falsone et al., 2004; Taipale et al., 2012). To provide further evidence for this, we assessed protein stability in cellular lysates by measuring solubility in a cellular thermal shift assay (CtTSA). We found that the majority (5 of 6) of mutant proteins with increased chaperone binding also exhibited decreased stability as measured by CtTSA (Figures S3A–S3D). In addition, computational predictions by the FoldX program (Schymkowitz et al., 2005) suggest that mutant proteins with increased binding to QCFs are likely to be significantly less stable than their WT counterpart (Figure S3D and Table S2B). Taken together, experimental and computational analyses
Figure 2. Most Disease Missense Mutations Do Not Impair Protein Folding or Stability

(A) Differential Z score distributions in LUMIER assay. Normalized differential Z scores are calculated as the difference in chaperone binding between all mutant/WT pairs expressed at detectable levels (n = 12,131). Non-expressed pairs serve as controls (n = 1,567).

(B) HSP90

(C) HSC70

(D) BAG2

(E) CHIP

(F) PSMD2

(G) GRP78

(H) GRP94

(I) Cytoplasmic ER

(J) AMPD3 VDR PITX2 GNAS CBS RPL11 NCF1 PAFAH1B1

WT W468R A320V C190W G33D WT R64R L54Q WT A366S R201C WT L154Q L456P P48L WT L20H WT R42Q T53A WT H277P S389R

α-HSP90β α-HSC70 α-FLAG

Differential interaction scores in LUMIER assay

(legend continued on next page)
suggest that mutant proteins with enhanced binding to QCFs have a destabilized protein structure.

Our quantitative survey of allele-specific interactions estimates that the majority of missense disease mutations do not dramatically impact protein structure or folding (Tables S1D and S2). Therefore, they may exert their deleterious effects through other mechanisms such as perturbation of molecular interactions.

**Disease Mutations and PPI Perturbations**

In principle, the effects of missense disease mutations on molecular interactions (Zhong et al., 2009), or “edgotype” (Sahni et al., 2013), could range from no apparent detectable change in interactions (“quasi-WT”), to specific loss of some interaction(s) (“edgetic”), to an apparent complete loss of interactions (“quasi-null”) (Figure 4A). To systematically characterize PPI perturbations associated with disease mutations and identify potential gain of interactions, we used the yeast two-hybrid (Y2H) interaction assay followed by a stringent validation assay. After autoactivator removal, we screened 2,449 mutant proteins and their 1,072 corresponding WT proteins for interactions with proteins encoded by the ~7,200 ORFs in the human ORFeome v1.1 (Rual et al., 2004). Mutant and WT proteins were then tested pair-wise against all partners found both in these Y2H screens and in our human interactome map HI-II-14 (Rolland et al., 2014) (Figure 1C). Altogether, we obtained interaction profiles for 460 mutant proteins and their 220 WT counterparts and found 521 perturbed interactions out of 1,316 PPIs (Table S3A).

To validate these results, we used the orthogonal in vivo Gaussia princeps luciferase protein complementation assay (GPCA) performed in human 293T cells (Cassonnet et al., 2011) (Table S3B). Unperturbed interactions were recovered at a rate statistically indistinguishable from that of a well-documented positive reference set (PRS), similar to the interactions of the WT alleles (Braun et al., 2009; Venkatesan et al., 2009). Perturbed interactions were recovered at a rate as low as a negative control “random reference set” (RRS) (Figures 4B and S4A), demonstrating the high quality of the identified perturbations induced by disease mutations.

To analyze global and topological characteristics of gene products with edgetic, quasi-null, or quasi-WT mutations, we used the human interactome map HI-II-14 (Rolland et al., 2014). According to the studied network properties (betweenness, k-core centrality, degree, closeness), the nodes (genes) examined in our edgotyping study appear unbiased, in that their topological properties are statistically indistinguishable from other genes in the network (Figures S4B–S4F). Interestingly, we found that the genes carrying edgetic mutations tend to be more central than either non-edgetic genes or the rest of the network (Table S4).

Out of a total of 197 mutations, corresponding to 89 WT proteins with two or more interaction partners, our interaction profiling identified 26% as quasi-null alleles, 31% edgetic and 43% quasi-WT (Figure 4C and Table S3C). We also analyzed...
disease mutations annotated by ClinVar (Landrum et al., 2014) and found the distribution of quasi-null, edgetic, and quasi-WT alleles was statistically indistinguishable from that of HGMD (Figure S4G). We only identified two mutations that conferred PPI gains, suggesting that gain of interactions may be a rare event in human disease.

Figure 4. Interaction Perturbation Profiles Distinguish Disease Mutations from Non-Disease Variants
(A) Three classes of PPI profiles (edgotypes) for mutations.
(B) Percentage of protein pairs recovered in GPCA at increasing score thresholds. Shading indicates SE of the proportion.
(C) Distribution of different edgotype classes for disease mutations.
(D and E) Differential LUMIER interaction scores among different edgotype classes, for binding to HSP90 (D) and HSC70 (E). p values by one-sided unpaired t test.
(F) Differential expression among different edgotype classes (ELISA log2 ratio of mutant over WT). QW: n = 75, E: n = 49, QN: n = 42. p values by one-sided Wilcoxon rank sum test.
(G) Distribution of different edgotype classes for non-disease variants.
(H) Increased binding to HSP90, HSC70, or either (Union) for non-disease (N) or disease (D) variant proteins. p values by one-sided Fisher’s exact test. Error bars indicate SE of the proportion. *p < 0.05.
See also Figures S4 and S5.
Protein Folding and Expression Levels of Edgetic Mutations

Differences between edgotype classes could be due to protein folding and/or relative expression levels. Quasi-null proteins associated significantly more with cytoplasmic, but not ER, chaperones, whereas edgetic and quasi-WT proteins did not significantly change their chaperone association (Figures 4D–4E, and S5A–S5E). Quasi-null proteins appeared to be poorly expressed, while edgetic and quasi-WT proteins were expressed at levels similar to those of their WT controls (Figure 4F). We validated several mutant-chaperone interactions and expression profiles by co-immunoprecipitation with endogenous chaperones, followed by western blot (Figure S5F). All tested quasi-null proteins exhibited more binding to HSP90 and HSC70, although they were expressed at lower levels than their WT controls. However, the edgetic TAT-P22OS protein and the quasi-WT NCF2-R395W protein did not show any detectable chaperone association. Among mutant proteins with no change in chaperone binding, edgetic (28%) and quasi-WT (57%) proteins comprised the majority, while quasi-null proteins comprised a significantly lower percentage (15%) (Figure S5G). Altogether, these results suggest that quasi-null proteins are more often unstable/misfolded and diminished in their steady-state expression levels. In contrast, edgetic and quasi-WT proteins likely exhibit normal folding and expression levels, further supporting the idea that they may cause disease through interaction perturbations or other mechanisms rather than simple loss of protein function.

Disease-Causing Mutations Versus Common Variants

Genome-wide association studies have identified hundreds of loci linked to particular disorders. However, these loci often contain several genes and multiple variants, making it challenging to distinguish causal mutations from non-pathogenic variants. We observed previously that among binary interactions found by WT proteins, disease-causing alleles were more likely to perturb interactions than non-disease variants (Rolland et al., 2014). We further investigated both disease-causing alleles from HGMD and common variants identified in healthy individuals from diverse geographical sites (1000 Genomes Project Consortium, 2012) (Table S1A) with respect to the edgetic character and chaperone binding of their protein products. Interaction profiling showed that only a small fraction of non-disease alleles lost interactions (8%, Figure 4G), a 7-fold reduction relative to disease mutations (57%; p = 1.7 × 10⁻³; Figure 4C). In addition, non-disease alleles on average did not alter chaperone association (Table S2A), a characteristic distinct from disease mutations annotated by HGMD (Figure 4H) or ClinVar (Figure S5H). Together, interaction perturbations can help distinguish disease-associated alleles from non-disease alleles.

To assess the predictive power of edgotyping to identify disease-causing mutations, we determined its precision and sensitivity in classifying an allele as causal based on interaction perturbation profiles. As a “gold standard” for causal alleles, we used a set of mutations annotated in HGMD as disease-causing (“DM” in Table S1A). As a negative control, we used a set of alleles most likely not associated with disease. We observed that 96% (105 of 109) of the alleles found to perturb interactions (E or QN) were disease-causing (Figure S6A). Conversely, 61% (105 of 172) of disease-causing mutations annotated by HGMD were interaction-perturbing (Figure S6B). Together, our prediction achieved a precision (96%) and sensitivity (61%) significantly higher than random expectation. It is possible that current incompleteness of interaction network maps might limit the power of edgotyping to properly classify disease-causing mutations. To evaluate this possibility, we performed a down-sampling analysis and found negligible effect on mutation classification over a broad range of network sizes (Figure S6C).

Edgetic Mutations and Interaction Interfaces

To explore edgotypes from a structural point-of-view, we assessed the possible impact of distinct classes of mutations on protein function using PolyPhen-2 analysis (Adzhubei et al., 2010). Interaction-perturbing mutations are significantly more often predicted to be deleterious than non-interaction-perturbing mutations (Figure 5A). We next investigated whether mutations from the different classes might differ in evolutionary conservation, based on the presumption that conservation of amino acid residues is a property that generally reflects functionality (1000 Genomes Project Consortium, 2012; Subramanian and Kumar, 2006; Sunyaev, 2012). The residues affected by interaction-perturbing mutations are significantly more conserved across species compared to non-interaction-perturbing mutations (Figure S6D). However, PolyPhen and conservation analysis could not distinguish between edgetic and quasi-null mutations within the interaction-perturbing group.

Given that structural domains often mediate protein interactions, different classes of mutation might vary in their locations relative to protein domains. Interaction-perturbing mutations are indeed significantly enriched within structural domains compared to non-interaction-perturbing alleles (Figure 5B and Table S1C). In addition to structural domains, intrinsically disordered regions and linear motifs could also play a role in mediating PPIs. However, we found interaction-perturbing disease alleles to be depleted in intrinsically disordered regions (Figure S6E), and occurring in linear motifs as frequently as non-perturbing alleles (Figure S6F). These results suggest that mutations perturbing PPIs are preferentially located within structural domains. Nevertheless, none of the above properties could reliably predict whether a mutation would give rise to an edgetic or quasi-null PPI effect.

We next investigated whether edgetic and quasi-null mutations differ in their physical location within three-dimensional protein structures (Zhong et al., 2009). Edgetic mutations are significantly more enriched in structurally exposed residues compared to quasi-null mutations (Figure 5C). Consistently, edgetic mutations do not tend to cause a change in hydrophobicity, a destabilizing feature that generally disrupts protein function (Balasubramanian et al., 2005), while quasi-null mutations often lead to a decrease in hydrophobicity (Figure S6G).

We also investigated whether or not edgetic mutations are more frequently located at an interface that supports interaction with a partner protein. Starting from all available structures of co-crystal complexes in the Protein Data Bank (PDB) involving a disease gene product, we determined the relative location of each mutated residue within these structures (Extended Experimental
Cell interactors exhibit a striking tendency to be expressed in disease-relevant tissues using RNA-seq compared gene expression patterns for perturbed and unperturbed partners in disease-relevant tissues. To test this, we hypothesized that protein interaction partners perturbed by edgetic mutations are likely to function together within the tissue and more likely to exhibit different edgotype classes of perturbed PPI profiles (Table S5C). In contrast to quasi-null mutations, edgetic mutations are significantly enriched at interaction interfaces identified from the corresponding co-crystal structures (Figure 5D). Notably, edgetic mutations also exhibit a significant tendency to reside at interaction interfaces with the perturbed partner, as compared to unperturbed partners or random controls (Figure 5E). These results suggest that edgetic mutations are preferentially located at PPI interfaces, perturbing the corresponding interaction.

Edgetic Mutations Perturb Interactions with Protein Partners Expressed in Disease-Relevant Tissues

We hypothesized that protein interaction partners perturbed by edgetic mutations are likely to function together within the tissue known to be affected by the relevant disease. To test this, we compared gene expression patterns for perturbed and unperturbed partners in disease-relevant tissues using RNA-seq data from the Illumina Human Body Map 2.0 project. Perturbed interactors exhibit a striking tendency to be expressed in disease-relevant tissues compared with unperturbed interactors or random genes (Figures 5F and S6H; Table S5B). These results indicate that disease mutations most often perturb interactions that are functionally relevant in the particular tissue(s) affected by a specific disease.

Distinct Interaction Perturbations May Underlie Diverse Disease Phenotypes

Our edgotyping model suggests that different mutations in the same gene may result in different, pleiotropic phenotypic outcomes through perturbation of distinct interactions (Figure 6A). To test this, we compared mutation edgotype classes and the resulting disease phenotypes. Among pleiotropic genes associated with two or more diseases, mutant alleles associated with different disease manifestations were more likely to exhibit different edgotype classes of perturbed PPI profiles (Table S5C).

This is exemplified by mutations in TPM3, which encodes slow muscle alpha-tropomyosin. Three TPM3 edgetic mutations (L100M, R168G, and R245G) are associated with fiber-type disproportion myopathy through an unknown mechanism (Adzhubei et al., 2010; Clarke et al., 2008) (Figure 6B). These edgetic mutations perturb five of the ten interaction partners of the WT gene product. The majority of perturbed partners are expressed in muscle, the tissue most relevant to this disease (Figure 6C). gene product. The majority of perturbed partners are expressed in muscle, the tissue most relevant to this disease (Figure 6C).
et al., 2012), could also be of disease relevance. In contrast to these edgetic mutations, the quasi-WT mutation M9R causes a different disease, nemaline myopathy. M9R might affect actin binding, thus leading to the formation of abnormal nemaline rods (Laing et al., 1995).

The possible disease relevance of our approach was further illustrated by edgetic mutations in the gene EFHC1, mutations in which can cause epilepsy. One perturbed partner, ZBED1, plays a role in a major cell proliferation pathway affected by EFHC1 knockouts (Yamashita et al., 2007), while another perturbed interactor, TCF4, is required for neuronal differentiation (Flora et al., 2007) (Figure 6D).

We next reasoned that mutations perturbing a greater number of interactions would be likely to have a larger impact on protein function, and hence result in more severe phenotypic effects. We used the age of disease onset as a proxy for severity and determined whether an increase in the fraction of interactions lost correlated with an increase in severity for each pair of mutations causing the same disease (as annotated by HGMD) (Figure 6E and Table S5D). We found that mutations perturbing more PPIs were associated with an earlier age of disease onset significantly more often than random expectation (Figure 6E). Although computational predictions based on PolyPhen-2 were able to distinguish between interaction-perturbing versus non-perturbing alleles (Figure 5A), they did not perform as well as our approach in predicting disease severity (Figure S6I). This limitation is consistent with the inability of PolyPhen-2 to distinguish between edgetic and quasi-null mutations (Figure 5A).

Protein-DNA Interactions
We hypothesized that mutations for which no PPI perturbation has yet been detected likely cause changes in other types of molecular interactions. As a proof-of-concept, we examined the effect of disease mutations on protein-DNA interactions (PDIs) between human transcription factors (TFs) (Reece-Hoyes et al., 2011a) and developmental enhancers (Fuxman Bass et al., 2011). Our hmoRFome1.1 mutant library contains 70 TF ORFs available, about half did not perturb any PDIs (Figure 7D). Interestingly, for ~80% of these we did identify PDI perturbations. For instance, mutations in the TGF-β-induced transcription factor TGF1 cause holoprosencephaly (Gripp et al., 2000). While the two mutant variants S28C and P63R are still able to bind their protein partners CTBP1 and CTBP2 (quasi-WT for PPI), both mutations completely abrogated the ability of TGF1 to bind any of the tested DNA targets (quasi-null for PDI) (Figure S7A). Clearly, integrating different types of molecular interactions will enhance our ability to understand specific mechanisms that underlie many genetic disorders.

To gain further insights into alternative molecular interaction perturbations, we computationally examined the effect of disease mutations on protein-chemical interactions (Reva et al., 2011). We found that the frequency with which disease mutations are at protein-chemical interfaces is significantly higher than that of non-disease variants (Figure S7B). In addition, disease mutations that perturb PDIs have no discernable tendency to locate at protein-chemical interfaces (Figure S7C), suggesting that protein-protein and protein-chemical interfaces do not tend to overlap. Interestingly, ~13% of PPI non-perturbing mutations are located at protein-chemical interfaces, supporting the conclusion that these mutations could cause disease through perturbation of alternative types of molecular interactions.

We combined computational predictions and interaction profiling to optimize our performance in disease mutation stratification. Although computational methods such as PolyPhen-2 could predict interaction-perturbing alleles as deleterious (Figure 5A), they fail to explain many disease-causing mutations, and misclassify them as “benign” (Figure S7D). Among these misclassified mutations, ~50% could be explained by molecular interaction perturbations (PCI, PPI, or PDI). For instance, the S140F mutation in PKP2 encoding the adhesion protein plakophilin leads to arrhythmogenic right ventricular dysplasia (Gerull et al., 2004). While PolyPhen-2 predicts S140F as benign, the S140F mutant exhibited increased binding to the chaperones HSC70 and BAG2, and lost all the PPIs of the WT protein (Table S7A). All together, existing computational methods alone fail to precisely predict disease causality. Examining different types of molecular interaction perturbations is critical for a full comprehension of disease-causing mutations in human.
Figure 6. Heterogeneous Genetic Mutations Give Rise to Diverse Disease Outcomes through Distinct Interaction Perturbations

(A) Schematic of pleiotropic disease outcomes resulting from distinct interaction patterns (edgotypes) caused by distinct mutations. Percentage of mutation pairs causing different diseases out of all pairs with different or the same edgotype classes is shown. n = 52. Error bars, SE of the proportion. p values by one-sided Fisher’s exact test.

(B) Example of edgotyping four disease mutations in the pleiotropic gene TPM3.
DISCUSSION

In this systematic characterization of mutations across various human Mendelian disorders, we have found surprisingly widespread disease-specific perturbations of macromolecular interactions. Approximately 60% of disease-associated missense mutations perturb PPIs, among which half result in complete loss of interactions, generally caused by protein misfolding and 

(C) Most perturbed partners of TPM3 are expressed in the disease-relevant tissue.

(D) Edgetic mutations in EFHC1 perturb epilepsy-related protein partners.

(E) Correlation between the fraction of PPI perturbation and age of onset for mutation pairs causing the same disease. p values by comparing the observed value to 100,000 random controls (n = 13; Extended Experimental Procedures).

See also Figure S6.
impaired expression, and the other half lead to edgetic perturbations. Importantly, different mutations in the same gene frequently result in different interaction perturbation profiles. This strongly suggests that the “edgotype” of a mutation represents a fundamental link between genotype and phenotype.

Our systematic edgotyping strategy provides a practical approach to classifying candidate disease alleles emerging from genome-wide association studies and from sporadic and somatic mutation sequencing approaches. Edgotyping achieves a high precision in identifying candidate disease-causing mutations based on the interaction perturbations relative to WT alleles (Figure S6A). However, the overall sensitivity of an edgotyping approach is compromised due to the false negative rate inherent to the assays used. We expect that a significant fraction of variants currently viewed as non-interaction-perturbing (quasi-WT) will eventually be proven to be edgetic and possibly cause disease. This circumstance likely arises from the incomplete nature of current human interactome network maps (Rolland et al., 2014). Nevertheless, because edgetic mutations cannot become quasi-WT or quasi-null even as interactome maps improve, our estimate of edgetic mutations already provides a reliable minimum lower bound for their frequency.

An alternative possibility is that quasi-WT mutations affect disease phenotypes through perturbation of different types of molecular interactions. Biological signaling is regulated at multiple levels, and various types of molecular interactions are involved (Sahni et al., 2013) as we have shown for PPI and PDI networks. In addition, protein-RNA (Lee et al., 2006) and protein-metabolite (Carpten et al., 2007) interactions have also been shown to be involved in disease. Perturbations of these alternative interaction networks will undoubtedly result in distinct disease consequences. One can envision that integration of additional types of interaction perturbation information with computational predictions will be necessary for a complete understanding of the cellular networks governing a particular disease state (Figure S7D). As a major benefit, perturbed interactions spotlight specific targets and pathways that are altered in a patient-specific context. This type of information could provide a much-needed guide in efforts to developing better diagnostic tools and more personalized medical treatments.

**EXPERIMENTAL PROCEDURES**

Using ORFs in the human ORFeome v8.1 collection as template, we PCR amplified the two DNA fragments flanking the mutations, followed by a fusion PCR to stitch the fragments together. The resulting fusion ORFs harboring the mutations were Gateway cloned into the Donor vector pDONR223 to derive Entry clones (Rual et al., 2004), which were subsequently verified by next-generation sequencing (Yang et al., 2011). Interaction with chaperones and other QCIs was performed using a quantitative LUMIER assay (Taipale et al., 2012; Taipale et al., 2014). All wild-type and mutant allele clones were transferred via Gateway recombination into a mammalian expression vector containing a C-terminal 3xFLAG-V5 tag. Stable HEK293T cell lines expressing luciferase-QCF fusion proteins were generated by lentiviral infection, and plasmids carrying wild-type and disease mutation alleles were transfected into the stable HEK293T lines (Taipale et al., 2012). Following capture of FLAG-tagged proteins, luminescence was measured to determine QCF-target interaction. Following luminescence measurement, FLAG-tagged mutant and wild-type proteins were detected as described (Taipale et al., 2012).

We performed a binary protein-protein interaction screen for all mutant and wild-type alleles as baits against ~7,200 human prey proteins (Rual et al., 2004). The identified interactions were combined with the known pairs cataloged by the human binary interaction dataset Hi-II-14 (Rolland et al., 2014). All first-pass pairs from the primary Y2H screens were subjected to pairwise testing in which all interactors of any allele of a gene were then tested against all alleles of that gene. The resulting verified protein-protein interaction profiles of disease mutants were compared with their wild-type counterparts. We validated perturbed and unperturbed interactions from mutation-mediated interaction perturbation data (“edgotyping” data) using an orthogonal in vivo method.

**SUPPLEMENTAL INFORMATION**

Supplemental Information includes Extended Experimental Procedures, seven figures, and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.04.013.

**AUTHOR CONTRIBUTIONS**


**ACKNOWLEDGMENTS**

We thank the members of the DFCI Center for Cancer Systems Biology (CCSB) for valuable discussions and acknowledge A.A. Chen, M. Koeva, and E. Guney for helpful suggestions. This work was supported by NHGRI (P50HG004233 to M.V., F.P.R. and A.L.B.; RC4HG006066 to M.V., T.H., D.E.H., K.S.-A., J.L.W., and S.L.; and R01HG001715 to M.V., D.E.H., and P.R.), NIH (GM082971 to A.J.M.W.), NSERC (CCF-1219007 to Y.X.), and the Krembil Foundation, a Canada Excellence Research Chair, an Ontario Research Fund–Research Excellence Award awarded to F.P.R. J.I.F.B. is supported by a Pew Latin American Fellowship. A.K. is supported by an EMBO Long-Term Fellowship. S.L. is an investigator...
REFERENCES


associate with the heat shock protein 90 chaperone and are destabilized following exposure to geldanamycins. Cancer Res. 65, 6401–6408.


EXTENDED EXPERIMENTAL PROCEDURES

Generation of the Human Disease Mutation ORFeome V1.1

Selection of Disease Mutations
The list of human disease mutations and associated phenotypic consequences was retrieved from the Human Gene Mutation Database (HGMD) public release version 2009. HGMD is a comprehensive repository of germ-line mutations in genes that are causal for, or are associated with, human inherited disease (Stenson et al., 2014). We mapped all the disease genes carrying single nucleotide mutations in HGMD to the human genome hg18 assembly using BLAT (Kent, 2002) with default settings, pursuing only those that could be unambiguously mapped. We mapped the mutation positions to the corresponding chromosomal coordinates in the reference genome. We then filtered the genes by their availability in our full-length sequence verified human ORFeome v8.1 collection (Yang et al., 2011), and obtained 16,381 mutations in 1,274 genes. For characterization of mutations across all diseases, we selected up to 4 mutations per gene, resulting in 3,443 mutations (Figure S1A). For genes with more than 4 mutations, the corresponding ORFs were sub-divided into four equally sized sections, and the mutation reported in the earliest publication was chosen in each section.

High-Throughput Mutagenesis
To generate disease mutations, we implemented an enhanced high-throughput site-directed mutagenesis pipeline described previously (Zhong et al., 2009) with modifications. Briefly, we first transferred the corresponding wild-type reference ORFs from their Entry clones in human ORFeome v8.1 into the pDEST-AD vector (AD domain from the yeast transcription factor Gal4) expressing the AD-ORF fusion proteins, followed by a 3-step PCR experiment. For a given disease mutation, PCR cloning consisted of two “primary PCRs” to generate gene fragments, and one “fusion PCR” to obtain the mutated ORF. For the primary PCRs, two universal primers, Tag1-AD and Tag2-TERM (sequences shown below), and two ORF-specific internal forward and reverse primers were employed (sequences shown in Table S1B). The two universal primers allowed the preservation of the attB sites on both ends of the ORF. The mutation-specific primers, MutF and MutR (Table S1B), encompassing the desired single nucleotide change, were designed to have an overlapping region of ~40 base pairs. The two ORF fragments flanking the mutation of a disease gene were amplified using the primer pair Tag1-AD and MutR, and the primer pair Tag2-TERM and MutF, respectively. For the fusion PCR, the two primary PCR fragments were fused together using the primer pair Tag1 and Tag2 (sequences shown below) to generate the single amino acid change (SAC) mutation allele. The final product was a full-length ORF harboring the desired SAC mutation. To create mutation Entry clones, all mutant ORFs were cloned into the Gateway donor vector, pDONR223, by a BP reaction followed by bacterial transformation into DH5α-T1R cells and selection for spectinomycin resistance. Four independent colonies per mutant ORF were isolated. For subsequent sequence confirmation, the inserts from the picked colonies were PCR amplified with KOD HotStart Polymerase (Novagen) using M13G-FOR and M13G-REV as primers (sequences shown below).

Primer sequences:

Tag1-AD: 5'-GGCAGACGCTCCTACCTACTCAGTGTTTGAACATCACTACAGGG-3'  
Tag2-TERM: 5'-CTGAGCTCTGACGCTACACGGTCCTCTGGC-3'  
Tag1: 5'-GGCAGACGCTCCTACCTACT-3'  
Tag2: 5'-CTGAGCTCTGACGCTACAC-3'  
M13G-FOR: 5'-CCCAGTCACGACGCTTTGAAGCGACCG-3'  
M13G-REV: 5'-GTGTCTACACCTCTCTGATGTTA-3'

Confirmation of Mutations Using Next-Generation Sequencing
To confirm expected mutations, the resulting PCR products were processed for 454 GS-FLX next-generation sequencing. First, the PCR products were pooled, ensuring that only one colony of one mutation from each gene was present in each pool. The pooled PCR products were purified using MinElute PCR Purification Kit (QIAGEN), and the DNA content was quantified by a NanoDrop spectrophotometer (Thermo Scientific). Different pools were multiplexed by ligation to unique adaptors. The samples were processed with, in order, GS Standard DNA Library Preparation kit, GS-FLX Standard emPCR kit (shotgun), GS-FLX Standard LR70 Sequencing kit and GS-FLX PicoTiterPlate kit (70 × 75), according to the manufacturer’s instructions (Roche Applied Science). Reads from the 454 next-generation sequencing run were assembled and aligned to the reference ORF sequences using 454 gsMapper program (Roche) with default settings. The clones that had a full-length coverage with only the single desired mutation were selected and consolidated. Sequence verification confirmed a total of 2,890 disease mutations (84% success rate) with single amino-acid changes (SAC) in 1,140 genes (hmORFeome1.1, Table S1A). The mutations in hmORFeome1.1 span a wide spectrum of human diseases and allow for systematic characterization of the functional effects of mutations.

Data Availability
All clones described in this study are publicly available. To receive the mutant clones described in this manuscript, clone requests must include the unique allele identifier numbers (Allele_ID) provided in Table S1A (e.g., “2_18118” as an example for a specific entry clone: A2M mutation C972Y). All clones will be made available through the auspices of the ORFeome Collaboration (http://orfeomecollaboration.org). Please contact David E. Hill (david_hill@dfci.harvard.edu) or Tong Hao (tong_hao@dfci.harvard.edu) with any questions regarding clone availability and clone distribution procedures.

Primer sequences:

Tag1-AD: 5'-GGCAGACGCTCCTACCTACTCAGTGTTTGAACATCACTACAGGG-3'  
Tag2-TERM: 5'-CTGAGCTCTGACGCTACACGGTCCTCTGGC-3'  
Tag1: 5'-GGCAGACGCTCCTACCTACT-3'  
Tag2: 5'-CTGAGCTCTGACGCTACAC-3'  
M13G-FOR: 5'-CCCAGTCACGACGCTTTGAAGCGACCG-3'  
M13G-REV: 5'-GTGTCTACACCTCTCTGATGTTA-3'
RNA Abundance Analysis
The RNA abundance data were downloaded in January 2015 from Human Protein Atlas (www.proteinatlas.org). To compare genes from different datasets, we used the RNA abundance levels from the HEK293 cell line. Three datasets were included for analysis: genes in the hmORFeome1.1 library, genes with disease-causing mutations in the HGMD database, and all human genes described in the Consensus CDS database (http://www.ncbi.nlm.nih.gov/CCDS/; downloaded October 2012). When multiple ORFs exist for the same gene, values were computed for each available ORF separately and then the averaged values were assessed. Gene IDs in each dataset were mapped to Ensembl gene IDs from the NCBI database (ftp://ftp.ncbi.nih.gov/gene/DATA/). Three RNA abundance levels “low,” “medium,” “high” were classified based on the following criteria: “0–10 FPKM,” “10–50 FPKM” and “>50 FPKM,” respectively. p values were determined by two-sided Fisher’s exact test.

GO Term Analysis
GO cellular component annotation data were downloaded from the UniProt-GOA database (Huntley et al., 2015), which combines high-quality electronic mappings and manual curation. Three datasets were included for analysis: genes in the hmORFeome1.1 library, genes with disease-causing mutations in the HGMD database, and all human genes described in the Consensus CDS database (http://www.ncbi.nlm.nih.gov/CCDS/; downloaded October 2012). Given that disease genes are likely to be more studied than other human genes (therefore having more GO terms), we resampled the HGMD and CCDS datasets with the same sample size as hmORFeome1.1 gene set to ensure the same GO annotation number distribution for the three datasets (bin width = 10). This procedure was repeated 100 times to obtain an average value. p values were determined by two-sided Fisher’s exact test.

Protein Domain Assignment and Mutation Mapping
We mapped Pfam domains (Pfam-A family only) to our set of genes using Hmmer version 3 (Finn et al., 2011) at an E-value threshold of $1 \times 10^{-5}$. The protein sequences were obtained by translating the cDNA sequences of the genes from each dataset. The presence of a Pfam domain and the corresponding range (start and end residues) were computed (Table S1C), and confirmed by the Pfam 27.0 database (March 2013, 14,831 families). For gene-level analysis, three datasets were included: genes in the hmORFeome1.1 library, genes with disease-causing mutations in the HGMD database, and all human genes described in the Consensus CDS database (http://www.ncbi.nlm.nih.gov/CCDS/; downloaded October 2012). p values were determined by two-sided Fisher’s exact test.

For mutation-level analysis, missense disease mutations and non-disease variants were each mapped relative to the range of the Pfam domains (inside a Pfam domain versus outside). We included only the alleles in genes that had at least one Pfam domain and at least one wild-type PPI. To assess the statistical significance of the differences in the fraction of mutations from different classes that reside within a Pfam domain, we employed a position shuffling approach, to better control for the variations in the domain content of proteins. To get an empirical p value, we randomly shuffled the positions of all mutations 100,000 times within their respective protein and calculated how often the resulting difference in fractions between two classes was as great as that observed in the real dataset.

Protein Intrinsic Disorder Analysis
For each residue affected by a genetic variant, we assessed the likelihood that the residue was located in an intrinsically disordered region of the protein (Table S1C). We subjected the protein sequences of the translated wild-type ORFs to the IUPred program (Dosztányi et al., 2005). Disorder was identified for the residues with a predicted probability higher than 0.5 in the protein sequence but region of the protein (Table S1C). We mapped Pfam domains (Pfam-A family only) to our set of genes using Hmmer version 3 (Finn et al., 2011) at an E-value threshold of $1 \times 10^{-5}$. The protein sequences were obtained by translating the cDNA sequences of the genes from each dataset. The presence of a Pfam domain and the corresponding range (start and end residues) were computed (Table S1C), and confirmed by the Pfam 27.0 database (March 2013, 14,831 families). For gene-level analysis, three datasets were included: genes in the hmORFeome1.1 library, genes with disease-causing mutations in the HGMD database, and all human genes described in the Consensus CDS database (http://www.ncbi.nlm.nih.gov/CCDS/; downloaded October 2012). p values were determined by two-sided Fisher’s exact test.

Linear Motif Analysis
The regular expressions of known eukaryotic linear motifs (ELMs) were downloaded from the ELM database (http://elm.eu.org/). The protein sequences were obtained by translating the cDNA sequences of the genes from each dataset. We filtered for matched sequence with a length $\geq 5$ residues in each protein. The number of matches to any motif from the ELM overlapping each mutation was computed. Motifs of the same type were not allowed to overlap but motifs of different types were allowed to overlap. Fisher’s exact test was used to determine whether certain classes of mutations were more likely than others to overlap one or more motifs.

Disease Mutation Annotations by HGMD
There are four different annotations for disease mutations in the HGMD database: disease-causing mutations (DM), disease-associated polymorphisms (DP), disease-associated polymorphisms with functional evidence (DFP) and functional polymorphisms (FP) (Stenson et al., 2014). DMs were reported to confer the associated clinical manifestations. These mutations were identified exclusively in patients stricken by the indicated disease but not in healthy individuals in the original literature report. In certain publications, the identified DMs were further substantiated with biochemistry and/or cell biology experiments as functional evidence for disease causality. DMs were therefore not likely to represent a polymorphism but were considered pathological mutations in human disease. The other three annotations included polymorphisms that were found in non-diseased individuals as well as diseased individuals. DP
mutations were reported to be significantly associated with disease (p < 0.05), but direct evidence of functional effect was not yet available. FP mutations were annotated based on their in vitro or in vivo functional effect, but these mutations had no disease association reported. DFP mutations were reported as disease-associated polymorphisms with additional supporting functional evidence.

**Protein-Chaperone Binding Assay**

**LUMIER Assay and High-Throughput Gateway Cloning**

We assessed mutant protein folding ability using chaperone binding as a proxy with a modified LUMIER assay (Taipale et al., 2012). We transferred all wild-type and mutant allele clones by Gateway recombination into a mammalian expression vector containing a C-terminal 3xFLAG-V5 tag. Inserts were then verified by restriction digestion; clones that did not produce the expected digestion pattern were omitted from further analysis. We rearrested the resulting 992 wild-type and 2,332 mutant clones such that wild-type and mutant alleles were in adjacent wells on the same plate. We then assayed all proteins for interaction with HSP90β (HSP90AB1), HSC70 (HSPA8), BAG2 (BAG2), CHIP (STUB1), PSMD2 (formerly known as RPN1), GRP78 (HSPA5), and GRP94 (HSP90B1). Stable cell lines expressing HSP90β, HSC70, BAG2, CHIP, and PSMD2 fused to Renilla luciferase had been previously described (Taipale et al., 2012; Taipale et al., 2014). For GRP78 and GRP94, Ophophrus luciferase (NanoLuc) (Heise et al., 2013) was inserted into the coding sequence after the signal peptide (GRP78: AEEE-NanoLuc-SGS-EEEDKK, GRP94: SVRA-NanoLuc-SGS-DDEVOD).

**Assessing Chaperone Binding with a Quantitative LUMIER Assay**

Stable HEK293T cell lines expressing luciferase-chaperone fusion proteins were generated by lentiviral infection (Taipale et al., 2012). Plasmids carrying wild-type and disease mutation alleles were transfected in 96-well plates into the stable cell lines using polyethyleneimine (Polysciences). At 48 hr post transfection, cells were washed in 1xPBS buffer and lysed in HENG buffer (50 mM HEPES-KOH [pH 7.9], 5% glycerol, 20 mM Na2MoO4, 150 mM NaCl, 2 mM EDTA, 0.5% Triton X-100 supplemented with protease inhibitors). The lysate was transferred to 384-well plates (Greiner) coated with anti-FLAG M2 antibody (Sigma-Aldrich), and incubated for three hours at 4°C. The plates were then washed again with HENG buffer, and the luminescence, as a measure of the level of chaperone interaction with the wild-type or mutated protein, was measured with an Envision plate reader (Perkin-Elmer) using Gaussia FLEX luciferase kit (New England Biolabs) or Nano-Glo Luciferase Assay System (Promega). Following luminescence measurement, horse-radish peroxidase (HRP)-conjugated anti-FLAG antibody in ELISA buffer (1xPBS, 2% goat serum, 5% Tween 20) was added to the wells. One hour later, the plates were washed in 1xPBS/0.05% Tween 20, and ELISA signal was read after adding chemiluminescent substrate (SuperSignal ELISA Pico, Thermo Scientific) as a semiquantitative measure of protein expression for both wild-type and mutant proteins.

**LUMIER Data Normalization and Scoring**

To adjust for plate variation, every plate contained the same set of positive and negative controls used for normalization and as a standard for ELISA signal. We applied a two-step normalization method. We processed the raw luminescence scores using a probabilistic mixture model to minimize spatial biases and batch effects (Taipale et al., 2012; Taipale et al., 2014). The observed luminescence is assumed to be the summation of background luminescence modulated by a spatial bias and actual luminescence from true interactions. Specifically, we processed the log-luminescence values with a Gaussian mixture model to correct for the spatial bias. This mixture model has three components: a background log-luminescence derived from a Normal distribution from true interactions. Specifically, we processed the log-luminescence values with a Gaussian mixture model to correct for the spatial bias. This mixture model has three components: a background log-luminescence derived from a Normal distribution with prior parameters that accounts for the differences between plate-specific noise mean values. 2. We established a Gaussian process model to correct for the spatial bias on a plate. 3. If the bait and prey do not interact, the observed log-luminescence value is a combination of the background noise and the spatial bias for the given well. 4. For each plate p of prey g, we also have a mixture parameter πgp, which is the fraction of positive interactions on that plate, and is derived from a Beta distribution with prior parameters a and b. Then for each well w on the plate, we have a binary variable zgw Derived from a Bernoulli distribution with the mixture parameter πgp. The spatial bias bgp for all wells on a plate p of prey g is calculated from the Normal distribution N(0,K).

In summary, our model is defined as follows:

\[ x_{gw} \sim (1 - z_{gw})N\left(b_{gw} + u_g, \sigma^2_g\right) + z_{gw}N\left(b_{gw} + u_g + u_e, \sigma^2_g + \sigma^2_e\right), \]

where \( \sigma^2_e \) is the variance of the background log-luminescence distribution for prey g, and \( u_e \) and \( \sigma^2_e \) are the parameters for the log-luminescence distribution from a significant interaction. The detailed model inference has been described previously (Taipale et al., 2014).

For ELISA scores, we used a linear rescaling method for normalization. For each plate, we computed the 10% and 95% quantile thresholds from raw ELISA scores. To reduce the noise introduced by outliers, we linearly rescaled the ELISA
scores on all plates, such that the 10% threshold was set to 0 and the 95% threshold was set to 1. For each well \( w \) on plate \( p \):

\[
ELISA_{\text{norm}}(p, w) = \frac{ELISA_{\text{raw}}(p, w) - ELISA_{\text{20,1}}(p)}{ELISA_{\text{20,95}}(p) - ELISA_{\text{20,1}}(p)}
\]

We used ELISA values as a way to filter out proteins that were not expressed at high enough levels. Increased prey/bait ratio could be due to increased luminescence or decreased bait expression. If the bait is already expressed at very low levels (close to the sensitivity cutoff), small changes in bait expression could inflate the interaction scores and introduce artifacts.

**Validation of LUMIER Data by Co-Immunoprecipitation**

Mutant and wild-type proteins were tagged with a 3xFLAG epitope and transfected into 293T cells. Their interaction with endogenous HSP90 or HSC70 was assayed by co-immunoprecipitation (co-IP) followed by western blotting. The expression levels of mutant and wild-type proteins were measured with an anti-FLAG antibody following IP.

**LUMIER Mutant Differential Analysis**

After data normalization, differential LUMIER Z scores were computed to assess the fold change in chaperone binding between the mutant and the corresponding wild-type alleles. Wild-type/mutant pairs were included in the analysis if both alleles had an ELISA score > 0.20. Mutant or wild-type proteins with an ELISA score < 0.20 were considered not expressed in a particular experiment, and hence removed from the analysis. Welch’s t-test with pre-calculated chaperone-specific variances from all plates was applied to compute the differential Z scores. All tested pairs with differential LUMIER Z scores are provided in Table S2A.

To evaluate the chaperone expression levels across wells, we consider all wild-type/mutant pairs based on the normalized ELISA scores. To estimate the false positive rate of differential chaperone association at different cutoffs, we used non-expressed wild-type/mutant pairs as a background set, and all seven chaperones/quality control factors were included. We calculated the differential Z score distribution of all wild-type/mutant pairs in which both alleles had an ELISA score < 0.10. In these cases, Z score differences should reflect experimental variation and noise. We then compared the distribution of differential Z scores in the non-expressed set and the expressed set, and set the cutoff such that less than 2.5% of non-expressed pairs would be scored as significant. Since multiple testing was corrected for this analysis, the 2.5% should be considered the final false positive rate for the union of all seven chaperones/quality control factors.

**Clustering**

Quality control factors were clustered based on their similarities in differential Z scores (mutant versus wild-type interactions) by hierarchical clustering (average linkage and centered Pearson correlation), as implemented in MeV v4.8.1 (Saeed et al., 2006). Only those mutant/wild-type pairs were included in which both alleles were detectably expressed (as measured by ELISA).

**Mutation Characterization by Pairwise Testing Protein-Protein Interactions**

To identify PPI perturbations caused by disease mutations, we performed two experiments. In the first experiment, a systematic screen was conducted to search for novel interactions with mutant and wild-type ORFs, and for possible mutant-specific gain of interactions. The screen was performed using the yeast two-hybrid (Y2H) assay against a collection of 7,200 proteins from human ORFeome v1.1 (Rual et al., 2004), and followed by a pairwise test to interrogate mutation-induced PPI alterations. In the second experiment, all the available binary PPIs for the disease genes in our hmORFeome1.1 collection were identified in the most recent human interactome HI-II-14 (Rolland et al., 2014). PPI profiles were compared between the disease mutants and the respective wild-type alleles by a pairwise test against their interactors using the Y2H assay.

**Bait and Prey Strains and Yeast Transformation**

All the disease mutant ORFs and their corresponding wild-type counterparts were transferred by Gateway LR reactions into the pDEST-DB (CEN-based) vector, expressing the yeast Gal4 DNA-binding domain fusion proteins (DB-ORF). The resulting LR products were transformed into bacterial competent cells DH5α-T1R, followed by selection for ampicillin resistance. The DB-ORF plasmids were then extracted and transformed into the haploid yeast Y8930, mating type MATα, with the genotype leu2-3,112 trp1-901 his3Δ200 ura3-52 gal4Δ gal80Δ GAL2::ADE2 GAL1::HIS3@LYS2 GAL7::lacZ@MET2cyh2R, and selected on synthetic complete (SC) agar media without leucine (SC-Leu) to generate Y2H bait strains (Dreze et al., 2010). The prey strains expressing the Gal4 activation domain fusion proteins (AD-ORF) were derived from the yeast Y8800, mating type MATα, with the same genotype as Y8930, and selected on SC media without tryptophan (SC-Trp) (Rolland et al., 2014).

**Identification and Elimination of Auto-activators**

Prior to interaction tests, DB-X auto-activators (Walhout and Vidal, 1999) that could activate the GAL1::HIS3 reporter gene expression in the absence of an AD-Y plasmid, were identified and eliminated as described (Dreze et al., 2010). Strains containing individual mutant or wild-type DB-X constructs were mated with Y8800 strain carrying an empty pDEST-AD vector. After diploid selection in liquid SC-Leu-Trp media, the yeast cells were spotted onto SC agar media lacking leucine, tryptophan, and histidine but supplemented with 1 mM 3-amino-1,2,4-triazole (3AT) (SC-Leu-Trp-His+3AT). Any DB-X strains that exhibited growth on the SC-Leu-Trp-His+3AT media were identified as auto-activators, and eliminated before PPI screening.
**Systematic PPI Screen**

We employed an enhanced high-throughput robotic first-pass PPI screen pipeline, as described (Rolland et al., 2014). Briefly, after auto-activator elimination, we consolidated DB-X yeast strains at the gene level by pooling together mutant and wild-type clones from the same gene. After overnight growth, individual strains expressing DB-X constructs were mated with AD-Y minipool libraries from human ORFeome v1.1 (~7,200 genes) (Rual et al., 2004), each containing 188 unique AD-Y strains. Following mating in rich yeast extract peptone dextrose (YPEPD) liquid medium for 24 hr at 30 °C, the cells were transferred into liquid SC-Leu-Trp media to enrich for diploid yeast. After incubation overnight at 30 °C, 5 μl of the culture was spotted onto SC-Leu-Trp-His + 1 mM 3AT screening plates to select for activation of GAL1::HIS3 reporter gene expression. To detect de novo auto-activators that might arise during the screen, the same culture was simultaneously spotted onto SC-Leu-His + 1 mM 3AT + 1 mg/l cycloheximide (CHX) media plates. On CHX-containing media, auto-activator strains lost by plasmid shuffling the AD-Y plasmid expressing the sensitivity marker cych25, and activated GAL1::HIS3 reporter expression by DB-X alone, exhibiting a growth phenotype. As previously described (Dreze et al., 2010), all screening plates were incubated at 30 °C for 3 days. Yeast cells that grew on SC-Leu-Trp-His+3AT media but not on SC-Leu-His+3AT+CHX media were considered as primary positives. Yeast cells that grew on SC-Leu-His+3AT+CHX media were de novo auto-activators and were therefore excluded.

**Identification of Interaction Sequence Tags**

Given that every DB-X strain was mated against a pool of 188 AD-Y strains, it is possible that more than one interaction could be identified per minipool. Therefore, four colonies were picked per growth spot from primary positives, and each inoculated separately into SC-Leu-Trp-His + 1 mM 3AT liquid media. Yeast cells were then lysed in lysis buffer containing 2.5 mg/ml zymolyase 20T (Seikagaku), and the lysate was diluted as a template for PCR amplification of the DB-X and AD-Y inserts using HiFi Platinum Taq polymerase (Life Technologies), as described (Dreze et al., 2010). The PCR products were sequenced by Sanger sequencing protocols, and the DB-X and AD-Y sequencing reads were mapped to human mutation ORFeome v1.1 and human ORFeome v1.1, respectively, using BLASTN with an E-value cutoff of 1 × 10⁻¹⁵. The best hits of pairs of DB-Y and AD-X reads were assembled into interaction sequence tags (ISTs), if both reads of the same positive colony passed the cutoff. If the AD-Y sequence of an IST could not be unambiguously assigned to a single ORF because of multiple isoforms for the same gene present in the human ORFeome collection, we provisionally assigned such ISTs to all possible ORFs for further verification.

**PPI Pairwise Test**

All first-pass pairs from the primary Y2H screen was subjected to a pairwise test, as described (Dreze et al., 2010). Briefly, individual DB-X (disease mutation or wild-type clone) and AD-Y (interaction partner) yeast strains were inoculated in liquid culture SC-Leu and SC-Trp, respectively. These yeast cells were then mated in liquid YEPD media for 24 hr. After mating, the yeast culture was transferred into SC-Leu-Trp liquid media, and incubated for another 24 hr to enrich for diploid cells. The diploid yeast cells were then robotically spotted onto both selective and control agar plates, SC-Leu-Trp-His+3AT and SC-Leu-His+3AT+CHX, respectively, and the plates were incubated at 30 °C for 3 days. PPIs were considered positive for activation of GAL1::HIS3 (i.e., growth to overcome 3AT inhibition) in a CHX-sensitive manner. All protein pairs were tested 4 times independently in a 384-format. The pairs that scored positive three or four times were classified as pairwise positive PPIs. PPI comparisons of mutant clones with their wild-type controls gave rise to mutation-mediated PPI perturbations. In another experiment, interactions derived from HI-II-14 interactome dataset (Rolland et al., 2014) for all the disease genes were subjected to the same pairwise test pipeline described above, to compare the PPI profiles between mutants and their wild-type counterparts.

**Protein-DNA Interaction Assay**

Enhanced yeast one-hybrid (eY1H) experiments were performed as described previously (Reece-Hoyes et al., 2011a; Reece-Hoyes et al., 2013). Briefly, enhancer DNA was amplified by PCR using human genomic DNA as template (Clontech Laboratories) and cloned into the pDONR-P4-P1R vector by Gateway BP reactions to generate entry clones. After sequence verification by PacBio sequencing (Yale Center for Genomic Analysis), the enhancers were Gateway transferred by LR reactions to the pMW#2 and pMW#3 Y1H destination vectors that carry the two reporter genes, HIS3 and LacZ, respectively (Reece-Hoyes et al., 2011a; Reece-Hoyes et al., 2011b). To generate the DNA bait yeast strain, the bait vectors were linearized and transformed into the haploid MATα type yeast strain Y1H-aS2 where the DNA baits were integrated into the yeast genome. The wild-type and mutant TF preys were transferred by Gateway LR reactions to the destination vector pDEST-AD-2m (Life Technologies) to generate TF-AD (activation domain of the yeast TF Gal4) fusion proteins. Fusion clones were subsequently transformed into the haploid MATα yeast strain Yα1867. The DNA bait and the TF prey strains were mated on YAPD agar plates for 1 day, followed by a diploid selection for 2 days on SC media without Ura and Trp (SC–Ura–Trp). The resulting diploid yeast cells were then plated on SC–Ura–Trp–His media containing both X-gal (160 mg/L) and 3AT (5 mM), and incubated for one week to assay for PDIs.

Only those strains in which the TF-enhancer interaction enabled the expression of both the HIS3 and LacZ reporters would show growth of blue colonies on the selection media containing 3AT (5 mM) and X-gal (160 mg/L). Images were processed and analyzed using MyBrid (Reece-Hoyes et al., 2013) followed by manual curation. This eY1H pipeline utilized a Singer Rotor robot platform to manipulate the arrayed collection of TFs, which facilitated the comparison between wild-type and mutant TFs by allowing simultaneous testing on the same array plate.

DNA binding domain (DBD) sequences were obtained from the UniProt database (UniProt Consortium, 2014), and mapped to TFs by BLASTP (NCBI). The positions of TF mutations relative to DBDs were then determined. For each mutant TF, we calculated the
percentage of PDIs lost relative to its wild-type counterpart. To compare the percentage of PDI loss for TF mutations within versus outside DBDs, we plotted histograms with 5 bins dividing different percentage ranges, and the five bins are 0-12.5; 12.5-37.5; 37.5-62.5; 62.5-87.5; 87.5-100. One-sided Wilcoxon rank sum test was used to calculate the statistical significance of the difference in the percentage of PDI loss between TF mutations within versus outside DBDs.

Selection of Major Chaperone Classes for LUMIER Assay
For the LUMIER assay, we selected two of the three major chaperone classes (Hsp70 and Hsp90) in the cell. We did not assay the third major chaperone machinery in mammalian cells, the TRIC/CCT complex, because it is unfortunately not amenable to LUMIER assay. All TRIC/CCT subunits have both their C and N termini buried inside the folding cage, and thus it is not possible to tag them with a soluble domain without compromising the complex’s function. Indeed, we are not aware of any functional TRIC/CCT fusion proteins that have been reported in the literature. However, this chaperone system collaborates tightly with both Hsp90 and Hsp70, and many of its clients such as WD40-domain proteins (Yam et al., 2008) are also substrates for Hsp90 and Hsp70 as well (Taipale et al., 2014). Thus, the client repertoire of this chaperone complex is likely to largely overlap with those of the other two classes.

Solvent Accessibility and Structural Analysis
For protein structural analysis, we first predicted the solvent accessibility for each residue by the following algorithm. We collected 6,125 non-redundant and non-homologous proteins with high-resolution crystal structures from Protein Data Bank (PDB) (Rose et al., 2013). DSSP program (Andersen et al., 2002) was used to compute solvent accessible area for each residue. In order to compare different residue types, we normalized the absolute solvent accessible area values into relative solvent accessibility (RSA). We then annotated all residues into three categories: “buried” (B) with RSA < 10%, “medium” (M) with 10% ≤ RSA < 42.5%, and “exposed” (E) with RSA > = 42.5%. We finally trained a machine-learning model, conditional neural fields (Källberg et al., 2012), on this annotated sequence dataset. For each residue, we also incorporated 7 residues in both flanking regions for training. This method achieved 77% cross-validation predictive accuracy on classifying a residue into “B, M, E” categories. Finally, using forward-backward algorithm, we assigned to each genetic variant a reliable structural property (B, M or E; Table S1C) based on the computed probability, and calculated the fraction of each structural property in different classes of variants.

Annotating the Mode of Inheritance
We retrieved the mode of inheritance information for each disease mutation from two databases: Online Mendelian Inheritance in Man (OMIM) (Amberger et al., 2011) and Universal Protein Resource (UniProt) (UniProt Consortium, 2014). We assigned to each mutation a dominant or recessive inheritance based on the gene in which it resides and the disease with which it is associated. Mutations in genes with different inheritance modes were removed. To ensure the high quality of our automatic annotations, we also manually curated ~100 publications reporting these mutations to ensure the accuracy (Table S1C). We calculated the fraction of each inheritance mode in different classes of disease mutations, and evaluated the statistical significance by one-sided Chi-square test.

Estimating the Functional Impact of Mutations by PolyPhen-2
For prediction of possible impact of an amino acid substitution on the structure and function of a human protein, we subjected mutated residues to Polymorphism Phenotyping v2 (PolyPhen-2) analysis (Adzhubei et al., 2010). PolyPhen-2 predicts the functional significance of a genetic variant using Naïve Bayes classifier trained by supervised machine-learning. We selected “HumVar” dataset-trained PolyPhen-2 models, and the HumVar dataset consisted of all disease-causing mutations from UniProtKB and non-disease variants (minor allele frequency > 1%) most likely not involved in disease. We calculated the Naïve Bayes posterior probability (PolyPhen-2 score, Table S1C) of each mutation to be damaging, and obtained from PolyPhen-2 a false positive rate (FPR) at which the mutation was mis-classified as damaging. We qualitatively classified each mutation as probably damaging, possibly damaging, or benign, based on the false positive rate (FPR) thresholds (10% / 20% for HumVar model) (Table S1C). Mutations with their PolyPhen-2 scores associated with estimated FPRs at or below 10% are predicted to be probably damaging, between 10% and 20% as possibly damaging, above 20% as benign. Otherwise, the prediction would return an “unknown” result. The distribution patterns of PolyPhen-2 scores for different classes of mutations were compared using one-sided Wilcoxon rank sum test, calculated by R programming using the stats package.

Cellular Thermal Shift Assay
293T cells on 6-well plates were transfected with 3xFLAG-V5-tagged wild-type and mutant constructs. Two days later, cells were washed once with 1xPBS and resuspended in 1 ml 1xPBS. Cells were lysed by three freeze-thaw cycles with liquid nitrogen. Soluble proteins were separated by centrifugation and eight aliquots of the soluble fraction were transferred into a PCR microplate. Lysates were incubated for 5 min in a temperature gradient (37°C to 61°C) in a BioRad C1000 thermal cycler (Actual temperatures in the gradient: 37.0, 38.4, 41.6, 46.2, 52.0, 56.8, 59.5, 61.0). After incubation, lysates were centrifuged for 30 min at 13,000 rpm in a micro-centrifuge at 4°C and the soluble fraction was analyzed by SDS-PAGE and western blotting with an anti-FLAG antibody (Abcam ab2493, 1:10,000 dilution). ECL signal was detected with MicroChemi 2.0 Imager (FroggaBio) and bands were quantitated with ImageJ.
Prediction of Protein Stability Change by FoldX

We used the FoldX force-field algorithm (Guerois et al., 2002; Schymkowitz et al., 2005) to calculate the change in free energy of unfolding (ΔΔG) for all mutations that could be mapped to a published crystal structure. The detailed procedure was the following: We translated each SNP mutation into an amino-acid substitution in the corresponding protein sequence. For each protein, we used Smith-Waterman sequence alignment algorithm to search its wild-type amino acid sequence against all PDB structures. We removed cases where the substituted amino acid position is missing or unstructured in the PDB file. We also removed proteins with low-quality structures. Then, we used the FoldX program in a three-step procedure. First, we applied the “repair” option in FoldX for each wild-type atomic structure, which sampled from a probability-based rotamer library to remove steric clashes and other structural errors and to reconstruct missing side chain atoms. Second, we used FoldX to in silico mutate amino acid side chains in the atomic structure while also exploring alternative conformations of the proximal side chains. For each mutant, we repeat such “mutation” five times from random restart. Third, the structural stability of the wild-type structure and each of the five mutant models was calculated as the Gibbs free energy (ΔG) of unfolding using the FOLDEF empirical energy function. The change between the average ΔG of five mutant models and ΔG of wild-type structure was computed as the final ΔΔG. We correlated the FoldX scores with the binding data for 7 quality control factors, and computed the statistical significance by one-sided Wilcoxon rank sum test.

Validation of Interaction Network Perturbations

In Vivo Gaussia princeps Protein Complementation Assay

We validated perturbed and unperturbed interactions with high-throughput Gateway-compatible in vivo GPCA. GPCA detects protein-protein interactions by fragment complementation of Gaussia princeps luciferase (Cassonnet et al., 2011), pSPICA-C1 and pSPICA-N2 expression vectors each contain one of the two fragments of the humanized Gaussia princeps luciferase (Gluc1 and Gluc2, respectively) (Cassonnet et al., 2011; Remy and Michnick, 2006). The Gluc1 luciferase fragment is linked to the C terminus of a tested protein, and the complementary Gluc2 fragment is linked to the N terminus of another tested protein. Both mammalian expression vectors carry the same human cytomegalovirus (CMV) promoter and are maintained as high copy number with the human virus SV40 replication origin. The tested protein pairs were subcloned into the pSPICA-C1 and pSPICA-N2 vectors by LR clonase-mediated Gateway reaction (Life Technologies), and then transformed into DH5α bacterial competent cells and selected in ampicillin-containing 2xYT liquid medium. Plasmid DNA was extracted using QIAGEN 96 Turbo kits (QIAGEN) on a BioRobot 8000 (QIAGEN), and DNA concentration was measured by a high-throughput NanoDrop 8000 spectrophotometer (Thermo Scientific). The identity and orientations of DNA constructs were verified by sequencing, using the forward primer pCiNeo-Univ (5’-CAGCTCTTTAAGGCTAGATAG-3’) and the reverse primer pCiNeo-Rev (5’-CACTGATTCTCTAGGTGTTTGTG-3’). One configuration was tested in this validation experiment: All the wild-type and disease mutation alleles were cloned into the Gateway vector pSPICA-C1, and their respective interactors into the vector pSPICA-N2.

Cell Culture and DNA Transfection

Human embryonic kidney (HEK) 293T cells transfected with SV40 large T antigen were maintained in Dulbecco’s Modified Eagle’s medium (DMEM-Cellmax), supplemented with 10% fetal calf serum, 100 units/ml penicillin and 100 µg/ml streptomycin at 37° C with 5% CO2. For experimentation, 293T cells were seeded at a concentration of 3.0 x 10⁴ cells per well in 96-well plates (catalog No. 655 083, Greiner Bio-One). After one day of growth, cells were co-transfected with 100 ng of each construct expressing complementary fragments of the Gaussia luciferase fused in frame with the tested protein pairs. DNA transfection was performed using linear polyethyleneimine “Max” (catalog No. 24765, Polysciences) with a stock solution (1 mg/ml [pH = 7.0]). The DNA/PEI ratio (mass:mass) was 1:3. One day after transfection, cells were gently washed in 100 µl of DPBS buffer containing Ca²⁺ and Mg²⁺ (GIBCO) and incubated for 20 min in 40 µl/well of Renilla Buffer (Renilla Luciferase Assay system, catalog No. E2820, Promega) at room temperature with light shaking. The luciferase enzymatic activity was immediately measured by a GloMax Discover System luminometer (catalog No. GM3000, Promega). The Renilla luminescence counting program was used with an integration time of 10 s after injection of 50 µl of Renilla luciferase assay reagent (catalog No. E2820, Promega).

GPCA Scoring and Data Analysis

Protein pairs tested were defined as valid if both partners were successfully cloned into expression vectors, and each had 100 ng of plasmid DNA added during transfection. The GPCA signal was normalized as follows: For each tested protein pair X-Y, the luminescence unit of cells co-transfected with pSPICA-C1-X and pSPICA-N2-Y was divided by the luminescence control unit either from cells co-transfected with pSPICA-C1-X and empty pSPICA-N2 vector, or from cells co-transfected with pSPICA-N2-Y and empty pSPICA-C1 vector. The GPCA score is defined as the mean of the log2 of the two normalized luminescence values. The average GPCA score of two independent repeats was considered the final interaction score for each protein pair (Table S3B). The protein pairs were tested together with the PRS and RRS pairs scrambled randomly across all the same experimental plates. A quantitative output was used to titrate the “threshold” signal to an acceptable range. The threshold was set such that any pair scoring above that threshold was considered “positive” and otherwise considered “negative.” The recovery rate that was defined as the number of positive pairs divided by the total tested pairs should be viewed as a function of the score threshold.

Network Property Analysis

To determine the extent to which mutation edgotypes relate to topological features of the network, we performed network property analysis. We analyzed all the genes (nodes) carrying edgetic, quasi-null and quasi-wild-type mutations. In order to have sufficient coverage, we combined the systematic protein-protein interaction network Hi-II-14 (Rolland et al., 2014) with all new interactions
obtained from this edgotyping study, yielding a largest connected component (LCC) of 4,157 nodes and 13,989 edges (interactions). In this integrated network, there are a total of 86 genes from the edgotyping study (“S”): 36 edgetic genes (“E,” defined as genes carrying at least one edgetic mutation), and 50 non-edgetic genes (“NE,” defined as genes carrying only quasi-null or quasi-wild-type mutations). We compared these genes to all the 2,686 genes with degree > = 2 (“A”) in the network. The following network properties were analyzed: betweenness centrality (the fraction of all shortest paths in the network that pass through the node), k-core centrality (the largest sub-network comprising nodes of degree at least k), degree (the number of interactions a node has in the network) and closeness centrality (the inverse value of the average distance to all other nodes in the network). Furball network visualization was generated by the Pajek software (de Nooy et al., 2005), using the built-in Kamada-Kawai layout. Beanplots were shown to compare different sets of genes. p values were determined by one-sided Wilcoxon rank sum test.

Selection of Non-Disease Variants and Comparison to Disease Mutations

To compare the interaction perturbation profiles between disease mutations and non-disease variants, we restricted the search for non-disease variants to the same set of genes where disease mutations were found. For these genes, missense alleles with a frequency of > 1% were obtained from the 1000 Genomes Project (1000 Genomes Project Consortium, 2012) within human populations from diverse geographical sites. Alleles with > 1% frequency are considered as non-disease variants (common variants), based on the definition from UCSC Genome Bioinformatics Center (http://genome.ucsc.edu/goldenPath/help/haplotypes.html). Similar to disease mutations, we selected up to 4 non-disease variant alleles per gene, and this resulted in a final list of 70 non-disease variants in 48 genes. We then mapped all these non-disease variants to our full-length sequence verified ORFs in the human ORFeome v8.1 collection (Yang et al., 2011). To generate Entry clones for the non-disease variants, we employed the same high-throughput site-directed mutagenesis pipeline as for disease mutations, using wild-type ORFs in ORFeome v8.1 as template with primers listed in Table S1B. All non-disease variant clones were subsequently verified by next-generation sequencing (Table S1A).

To profile PPIs, the non-disease variant ORFs were Gateway transferred from the Entry vector into the pDEST-DB vector, and transformed into the yeast strain Y8930. For comparative purposes, all available non-disease variants, disease mutations and wild-type controls were included in the same experiment, using the enhanced Y2H pipeline described earlier. The results of interaction profiles for all non-disease variants and their edgotypes are listed in Table S3C. Two-sided Fisher’s exact 2 × 3 extension test was used to calculate the statistical significance between the edgotype distributions of disease mutations versus non-disease variants.

Analyses Using Disease Mutations Annotated by ClinVar

The ClinVar database (Landrum et al., 2014) was downloaded from NCBI. Only single amino acid change mutations were used for the filtering. We subjected all edgetic, quasi-null and quasi-wild-type mutations to ClinVar database, and obtained a subset annotated as pathogenic in ClinVar as well. We compared the edgotype distribution and chaperone binding for disease mutations annotated by HGMD versus those annotated by ClinVar.

Estimating the Predictive Power of Edgotyping for Disease-Causing Mutations

As a positive control for causative alleles, we used a set of mutations from HGMD that had been annotated as disease-causing (marked as “DM” in HGMD). As a negative control, we used a set of non-disease variants most likely not to be associated with disease. To assess if interaction-perturbing alleles (i.e., edgetic or quasi-null) are more causative compared to non-interaction-perturbing variants, we calculated the sensitivity, specificity, precision, and overall accuracy for our edgotyping-based prediction performance. These measures were calculated as follows:

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \\
\text{Specificity} = \frac{TN}{TN + FP} \\
\text{Precision} = \frac{TP}{TP + FP} \\
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

105 out of 109 (96%) alleles that had an edge-perturbing (i.e., edgetic or quasi-null) edgotype were causative (\(TP = 105, FP = 4\)), while only 67 out of 110 (61%) alleles with quasi-null-type edgotype were causative (\(TN = 43, FN = 67\)). Because non-disease variants and disease mutations had different sample sizes, we also calculated a balanced measure of classification performance, Matthews correlation coefficient (MCC), defined as:

\[
\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]

A Matthews correlation coefficient of +1 represents a perfect prediction, 0 indicates no better than random prediction, and −1 corresponds to a complete disagreement of the predictions. To further evaluate the predictive power of edgotyping, we randomly...
shuffled edgotypes 100,000 times for all the disease mutations and non-disease variants in our dataset. For each randomization, we kept the proportions of edgotypes constant. Then, we calculated the precision for disease-causing prediction in each randomized network, and then derived a distribution curve of precision calculations for all the shuffled networks. Statistical significance was calculated by comparing the observed empirical value to the 100,000 randomizations.

**Down-Sampling Analysis to Estimate the Effect of Network Completeness**

We randomly removed increasing fractions of interactions from the wild-type reference interaction network, and re-classified the edgotypes of each allele based on the remaining interactions. Genes with fewer than two remaining interactions were not considered in the analysis. We calculated the sensitivity, specificity, precision, and overall accuracy for the classification performance as a function of percentage of random interaction loss.

**Evolutionary Conservation Analysis**

The sequences of the wild-type ORFs in human ORFeome v8.1 corresponding to each genetic variant allele were translated to protein sequences. For each residue affected by a genetic variant, we estimated the evolutionary conservation levels by using PSI-BLAST to search the NCBI non-redundant sequence database with a 90% sequence identity cutoff. In the PSI-BLAST search, we included all 147 species with reference proteomes (http://www.ebi.ac.uk/reference_proteomes) in the database to cover both close and remote phylogenetic relationships, and ran the analysis with 5 iterations with an E value threshold of 0.001. We defined the conservation index for a residue as the fraction of related sequences in the multiple sequence alignment with matched residues (Table S1C). To compare the conservation levels for residues affected by different classes of genetic variants, we performed a one-sided Wilcoxon rank sum test.

**Analysis on Amino Acid Hydrophobicity Change**

We estimated the hydrophobicity as the propensity of residing in a solvent-excluded (buried) part of the protein by calculating the solvent-accessible surface area (SASA) of each residue over 2,241 representative structures from the Protein Data Bank (PDB), with the MSMS program (Sanner et al., 1996). We ranked each of the 20 residues by the ratio of the median SASA divided by the 95th percentile SASA, representing the maximal SASA for that residue. This resulted in the following ranking of hydrophobicity scores, CIVLAFMWGYTSHPNDRQEK. The residues that rank higher on this scale represent the ones that are more likely to be water-excluded (hydrophobic) in the context of real proteins. For each mutation, we calculated the absolute difference between the wild-type residue and the mutated residue by their rank on the hydrophobicity scale. Therefore, a positive hydrophobicity change score indicates a change from less to more hydrophobic, whereas a negative change score reflects a change toward low hydrophobicity. To assess the statistical significance, the fraction associated with a negative hydrophobicity change was compared for mutations of different classes, using one-sided Fisher’s exact test.

**PPI Interface Calculation**

Three-dimensional co-crystal structural data for PPIs was retrieved from the Interactome3D database (Mosca et al., 2013). All the protein structural files used in this analysis were downloaded from the Protein Data Bank (PDB, http://www.rcsb.org/pdb) (Rose et al., 2013) as of March 2014. To map human disease genes to PDB structures, we translated the full-length wild-type ORF sequences from human ORFeome v8.1 into protein sequences. Each disease protein sequence was then aligned with the most homologous PDB sequence using ClustalW (Larkin et al., 2007). Alignments with higher than 70% sequence identity over the entire length of the PDB sequences were retained for further analysis.

We took all available co-crystal structures of protein complexes in PDB. The structures with the longest sequence or the lowest E value were selected for each interacting pair. For PPI interface analysis, the mutated residues were first mapped onto the available structures by using Mechismo (http://mechismo.russelllab.org/), ProInDB (PROTein-protein INterface residues Data Base), and PDBePISA (Proteins, Interfaces, Surfaces and Assemblies) (Krissinel and Henrick, 2007) servers. Mutations were considered to reside at interaction interfaces of co-crystal structures if at least one of the following contacts between the mutated residue and a residue on the interactors was observed at the given distance:

1. The distance between alpha-carbon atoms at \( \leq 11 \, \text{Å} \);
2. The distance between the centers of the atoms at \( \leq 10 \, \text{Å} \); or
3. The distance between Van der Waals surfaces at \( \leq 5.5 \, \text{Å} \).

All the interfacial residues calculated by these servers were confirmed by mapping to protein structures manually by using PyMol Software (http://www.pymol.org). We analyzed over 100 protein interactions based on co-crystal structures for 70 disease mutations (corresponding to 29 genes) (Table SSA). The interface between two protein chains was generated by PyMol with a dASA (difference in accessible surface area between complex and single chains) cutoff of 0.75 square angstroms. Therefore, residues whose dASAs from the complex to a single chain were greater than 0.75 square angstroms were confirmed as interfacial residues. The relative solvent accessible surface area (rASA) was calculated by NACCESS (Lee and Richards, 1971) (http://www.bioinf.manchester.ac.uk/naccess) to determine if a protein residue was a surface residue or not (i.e., structurally buried). One-sided Fisher’s exact test was used to compare the distribution patterns of different classes of mutations at the structural level.
To investigate if the edgetic mutations are located on interfaces involving perturbed partners versus unperturbed partners, we computed the fractions of mutations observed at interfacial regions for both the perturbed group and the unperturbed group, as compared to a random control expected by chance. The random control was performed as follows. A similar number of disease mutations were randomly selected as controls based on available co-crystal structures. Mutation interface mapping was performed in the same fashion as edgetic mutations. To control for the difference in structural locations for mutations among different edgotype classes, we filtered for random mutations that do not belong to any edgotype class (their edgotypes were not determined in this study), so they are not biased toward any particular edgotype class. One-sided Fisher’s exact test was used to calculate the significant difference in the factions of interfacial mutations between the two groups. All the statistical analyses were done using the “stats” package in R.

### Gene Expression Analysis in Disease Relevant Tissues

For each edgetic mutation, we partitioned the interaction partners into two categories, mutation-perturbed and mutation-unperturbed partners. We then assessed whether or not each one of the perturbed and unperturbed interaction partners is expressed in the disease-relevant tissue. We retrieved the RNA-seq expression data of the Illumina Human Body Map 2.0 Project (GEO accession: GSE30611) from the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra; Study ERP000546; Runs ERR030888-ERR030903). This dataset covered 16 tissues: adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscle (muscle), testis, thyroid, white blood cells (WBC). For disease tissue expression analysis, all 75 base-pair RNA-seq reads were mapped to Human ORFeome v5.1 ORFs with Bowtie (Langmead and Salzberg, 2012), using default settings. Gene expression levels for each gene in each tissue were then defined as the log2 transformations of the upper quartile-normalized gene read counts for the tissue, as recommended previously (Bullard et al., 2010). Only genes with normalized expression levels greater than −2.5 in at least one tissue and lower than −1 in at least one tissue were considered. This ensured that genes considered were expressed above the background levels in at least one tissue but were not constitutively highly expressed. To assign a call of “expressed” or “not expressed” for each gene in each tissue, while at the same time controlling for differences in the physiological expression levels of different genes, we selected as the expression threshold for each gene the expression level in the tissue with the upper-quartile highest expression. Each disease, when possible, was manually assigned to a single tissue, and for each mutation, the interaction partners (both perturbed and unperturbed) were classified either as expressed or not expressed in the disease-relevant tissue. As a control, 30 randomly selected ORFs from the Human ORFeome were also classified as expressed or not expressed in the disease-relevant tissue for each mutation. One-sided Fisher’s exact test was used to determine whether perturbed partners were more likely to be expressed in the disease-relevant tissue than unperturbed partners and randomly selected genes.

### Assessment of Edgotype-Phenotype Relationships

To assess if different interaction perturbations could account for distinct disease phenotypic consequences, we first identified all pleiotropic genes with mutations (not restricted to the mutations characterized in this study) involved in more than one disease as annotated by the Human Gene Mutation Database (HGMD). We then filtered for disease-causing mutations, annotated as DM by HGMD database, in these pleiotropic genes. To examine the edgotype-phenotype correlation, we restricted to mutations in genes with a node degree larger than one (i.e., more than one protein interaction), and analyzed if a pair of mutations with different edgotypes exhibited different disease phenotypes. Statistical enrichment for mutation pairs with edgotype-phenotype correlations was calculated by one-sided Fisher’s exact test.

### Disease Severity and Age of Onset Analysis

We curated from the literature information about age of disease onset for genetic mutations as a surrogate for disease severity. We included disease-causing mutations, annotated as DM by HGMD, that exhibit dominant and homozygous recessive inheritance for this analysis. We were able to retrieve the age of onset information for a total of 58 alleles of 23 genes. To limit the effect of stochastic variation in ages of onset, age values of individual patients were normalized into 15 human developmental stages (from 1 to 15) as defined (Kang et al., 2011). We reasoned that a higher degree of interaction loss induced by mutations could lead to a more severe impact on the human molecular interaction network, and thus to an earlier age of onset. To test this hypothesis, we compared pairs of alleles that were associated with the same disease and examined the correlation between the degree of interaction loss, and the age of onset. For each pair of disease mutations of the same gene, whenever the comparison is possible (i.e., not a tie), we compared both the degree of interaction loss and the age of disease onset (Table SSD). Statistical significance was evaluated by a permutation test where the ages of onset (developmental stages) were randomly re-assigned among the alleles of the same gene. We ran the permutation 100,000 times to obtain a random distribution curve, and the p values was calculated by comparing the observed number of concordant cases to the random expectations.

### Protein-Chemical Interface Prediction

Data for protein-chemical interactions was obtained from MutationAssessor web-server (Reva et al., 2011). Only mutations in disease genes with degree larger than one were counted when comparing PPI non-perturbing and PPI perturbing mutations.
SUPPLEMENTAL REFERENCES


Figure S1. Generation of Human Mutation ORFeome Version 1.1, Related to Figure 1

(A) Systematic selection of disease mutations and high-throughput generation of hmORFeome1.1.

(B) Comparison of RNA abundance levels in the HEK293 cell line for the disease genes in this study, all the disease genes in the HGMD database, and all human genes (Extended Experimental Procedures). This study: n = 1,040, HGMD: n = 3,099, All: n = 16,724.

(C) Comparison of GO cellular component annotations. This study: n = 1,139, HGMD: n = 3,335, All: n = 18,339.

(D) Comparison of the fraction of genes with Pfam domains. This study: n = 1,140, HGMD: n = 3,296, All: n = 18,482.

(E) Percentage of mutations in disordered regions, comparing disease mutations in this study versus all missense disease mutations in HGMD (2009 version). This study: n = 2,890, HGMD: n = 35,490.

(F) Percentage of mutations in Pfam domains. Only genes with at least one Pfam domain are included for analysis. This study: n = 2,749, HGMD: n = 25,798.

(G) Percentage of mutations in linear motifs. This study: n = 2,890, HGMD: n = 35,490.

Error bars from (B) to (G) indicate SEs of the proportion. p values from (B) to (G) by two-sided Fisher’s exact test (Extended Experimental Procedures).
Figure S2. Systematic Characterization of Disease Mutant Proteins for Their Interaction with Cellular Quality Control Factors, Related to Figures 2 and 3

(A and B) Cumulative distribution of differential luminescence Z scores in all quality control factor interaction assays with mutant (Mut) proteins and their wild-type (WT) counterparts. Analysis included only pairs in which both alleles were expressed at detectable levels (orange), as assessed by ELISA. The background set consisted of mutant/wild-type pairs where neither allele was expressed (blue). This control set was used to establish a cutoff for true interactions, such that less than 2.5% of the “noise” could be scored as positive (A). At this cutoff, a subset of mutant alleles exhibit increased chaperone binding (A). In contrast, decreased differential Z scores are indistinguishable from the background control set (B).

(C) Scatter plot comparing mutant allele expression levels with wild-type allele expression levels by semiquantitative ELISA assay after transfection. Correlation is measured by Pearson coefficient of determination, $R^2$.

(D) Steady-state protein expression levels of recessive versus dominant disease alleles. $p$ values by one-sided Wilcoxon rank sum test.

(E) Percentage of recessive mutations that exhibit increased chaperone binding. $p$ values by one-sided Chi-square test between chaperone-binding versus non-binding mutants. Dashed line represents the mean of all mutants.

(F) PolyPhen-2 score distribution for mutation alleles that exhibit increased chaperone binding. The higher the score, the more deleterious the mutation is predicted to be. $p$ values by one-sided Wilcoxon rank sum test between chaperone-binding versus non-binding mutants. Dashed line represents the median of all mutants.

“All” from (D) to (F) indicate all tested mutants. For $n$ values, see Table S7B. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. 
Figure S3. Mutant Variants that Associate with Chaperones Are Less Stable, Related to Figures 2 and 3

(A–D) 3xFLAG-tagged wild-type and mutant variants for AMPD3 (A), VDR (B), CBS (C), and PAFAH1B1 (D) were transfected into 293T cells. Equal aliquots of cellular lysates were subjected to the indicated temperature gradient, after which soluble proteins were separated from the denatured fraction by centrifugation and analyzed by SDS-PAGE and western blotting with an anti-FLAG antibody. White arrowheads indicate a cross-reacting (cr) band in the AMPD3 blot. Quantitation of band intensity is shown on right. The interaction status of each mutant variant with chaperones, as measured by LUMIER and co-IP (Figure 2J), is indicated.
Figure S4. Systematic Characterization of Protein-Protein Interaction Perturbations by Disease Mutations, Related to Figure 4

(A) High reproducibility between two independent biological repeats in GPCA. Correlation is measured by Pearson coefficient of determination, $R^2$.

(B) Beanplot of the betweenness centrality values (gene-level analysis). Betweenness centrality is defined as the fraction of all shortest paths in the network that pass through the node. Values are in a log scale, and shifted by $\varepsilon = 0.001$ for clarity. A: all the genes with degree $\geq 2$, $n = 2686$; E: genes having edgetic mutation(s), $n = 36$; NE: genes having only QN or QW mutation(s), $n = 50$; S: the full edgotyping study (namely E+NE), $n = 86$.

(C) Beanplot of the k-core centrality values (gene-level analysis). K-core centrality is defined as the largest sub-network comprising nodes of degree at least $k$, and k-core centrality is the largest $k$ value, for which the node is found in a k-core.

(D) Beanplot of the degree values (gene-level analysis). Degree is defined as the number of interactions a node has in the network. Values are in a log scale.

(E) Beanplot of the closeness centrality values (gene-level analysis). Closeness centrality is defined as the inverse value of the average distance to all other nodes in the network. The shown values are normalized to be between 0 and 1.

(F) Visualization of the analyzed network, indicating the location of the edgetic (blue) and non-edgetic (red) nodes. Larger node size indicates larger betweenness centrality values.

(G) Comparison of edgotype class distributions among the disease mutations annotated by HGMD versus ClinVar. For ClinVar, QW: $n = 30$, E: $n = 25$, QN: $n = 24$. $p$ values from (B) to (E) by one-sided Wilcoxon rank sum test. $p$ values in (G) by Fisher’s exact 2x3 extension test.
Figure S5. PPI Quasi-Null Disease Mutants Are Likely to Be Unstable and Require Chaperone Assistance, Related to Figures 4 and 5

(A–E) Boxplots comparing differential interaction Z scores among mutant proteins from different edgotype classes, for binding to BAG2 (A), CHIP (B), PSMD2 (C), GRP78 (D) and GRP94 (E). Mutation edgotype classes: quasi-wild-type (QW), edgetic (E), quasi-null (QN). The middle horizontal bars indicate median. The top and bottom of the box indicate 25th and 75th percentiles, and whiskers indicate 10th and 90th percentiles. p values by one-sided unpaired t test with equal variances. 

(F) Co-immunoprecipitation (co-IP) blots. Mutant and wild-type proteins were tagged with a 3xFLAG epitope and transfected into 293T cells. Mutants in red, blue, and purple represent QN, E and QW edgotype classes, respectively. Their interaction with endogenous HSP90 (top) or HSC70 (middle) was assayed by co-IP. The expression levels of mutant and wild-type proteins (bottom) were measured with an anti-FLAG antibody following IP. The differential chaperone interaction scores for mutant proteins (mutant over wild-type) in the LUMIER assay are shown below the blots.

(G) Distribution of edgotype classes for mutants with increased chaperone binding, and for mutants with no change in chaperone binding. Error bars indicate SEs of the proportion. Enrichment p values is determined by Fisher’s exact 2x3 extension test.

(H) Comparison of PCI for non-disease variant (N) and disease mutant (D) proteins with chaperones HSP90 and HSC70 by LUMIER. D, disease mutants annotated by ClinVar; N, non-disease variants. “Union” represents variants with increased binding to either HSP90 or HSC70. p values by one-sided Fisher’s exact test. Error bars indicate SEs of the proportion. For more n values, see Table S7B. *p < 0.05.
Figure S6. Relationship between Interaction Perturbations and Disease Phenotypes, Related to Figures 5 and 6

(A and B) Estimate of the predictive power of edgotyping for disease-causing variants. Known disease-causing mutations are from HGMD. Putative non disease-causing variants are from 1000 Genomes Project. Precision (A) and sensitivity (B) are calculated and statistical p values is derived from comparing the observed value (black arrow) to 100,000 randomized controls (n = 207; red curve, see Extended Experimental Procedures for details).

(C) Down-sampling analysis is performed by randomly removing edges from the wild-type interaction network, and the sensitivity, specificity, accuracy, precision and Matthews correlation coefficient (MCC) are measured at each point (see Extended Experimental Procedures). These measurements remain steady over a wide range of network sizes.

(D) Comparison of the mutated residues with respect to the conservation levels throughout evolution. 147 diverse species were included for PSI-BLAST analysis and the conservation index was defined as the probability of a residue being conserved across evolutionary species. p values are determined by one-sided Wilcoxon rank sum test.

(E) Percentage of mutations in intrinsically disordered regions.

(F) Percentage of mutations in linear motifs.

(G) Percentage of mutations with an amino acid change from hydrophobicity to hydrophilicity.

(H) The enrichment of disease genes expressed in the relevant tissues. As a negative control, genes were randomly selected from the expression dataset (see Extended Experimental Procedures).

(I) A poor correlation between high PolyPhen-2 score (predicted to be more deleterious) and early age of disease onset (relatively high severity) for pairs of mutations causing the same disease. A mutation pair with such correlation is defined as “concordant.” Statistical significance is assessed from comparing the observed value (black arrow) to 100,000 randomized permutation controls (red curve, Extended Experimental Procedures).

For more n values, see Table S7(B). Error bars (E-H) indicate SEs of the proportion. Enrichment p values (E-H) are determined by one-sided Fisher’s exact test.
Figure S7. Integration of Various Types of Molecular Interaction Profiling and Comparison with PolyPhen, Related to Figure 7

(A) An example of disease mutations in the gene TGIF1, showing the wild-type PPI profiles, but altered PDI profiles.

(B) Percentage of disease mutations and non-disease variants occurring at protein-chemical interfaces (Extended Experimental Procedures). Error bars indicate SEs of the proportion. p values by one-sided Fisher’s exact test. Disease mutations: n = 2,890, Non-disease variants: n = 70.

(C) Percentage of PPI non-perturbing and perturbing disease mutations occurring at protein-chemical interfaces. Error bars indicate SEs of the proportion. p values by one-sided Fisher’s exact test. Perturbing: n = 113, Non-perturbing: n = 84.

(D) Heat map showing the integration of various types of molecular interaction profiling with PolyPhen prediction. In the plot, each horizontal line represents an allele ordered by PolyPhen-2 score (1st column), with the highest score on top. Red means probably or possibly damaging, while blue means benign (classified by PolyPhen-2). Columns 2-8: allele binding data for seven quality control factors. Red color denotes increased binding than the corresponding wild-type counterpart, otherwise represented in blue. Columns 9 and 10 show PPI and PDI profiles, respectively. Red color denotes alleles perturbing interactions, otherwise represented in blue. In the case of PDI, gray color means alleles not tested. Details can be found in Table S7A and Extended Experimental Procedures.