Three-dimensional reconstruction of protein networks provides insight into human genetic disease

Xiujuan Wang[1,2,5], Xiaomu Wei[2,3,5], Bram Thijssen[4,5], Jishnu Das[1,2,5], Steven M Lipkin[3] & Haiyuan Yu[1,2]

To better understand the molecular mechanisms and genetic basis of human disease, we systematically examine relationships between 3,949 genes, 62,663 mutations and 3,453 associated disorders by generating a three-dimensional, structurally resolved human interactome. This network consists of 4,222 high-quality binary protein-protein interactions with their atomic-resolution interfaces. We find that in-frame mutations (missense point mutations and in-frame insertions and deletions) are enriched on the interaction interfaces of proteins associated with the corresponding disorders, and that the disease specificity for different mutations of the same gene can be explained by their location within an interface. We also predict 292 candidate genes for 694 unknown disease-to-gene associations with proposed molecular mechanism hypotheses. This work indicates that knowledge of how in-frame disease mutations alter specific interactions is critical to understanding pathogenesis. Structurally resolved interaction networks should be valuable tools for interpreting the wealth of data being generated by large-scale structural genomics and disease association studies.

Over the past few decades, a tremendous amount of resources and effort have been invested in mapping human disease loci genetically and later physically[1]. Since the completion of the human genome sequence, especially with advances in genome-wide association studies and ongoing cancer genome sequencing projects, an impressive list of disease-associated genes and their mutations have been produced[2]. However, it has rarely been possible to translate this wealth of information on individual mutations and their association with disease into biological or therapeutic insights[3]. Most of the drugs approved by the US Food and Drug Administration today are palliative[4]—they merely treat symptoms, rather than targeting specific genes or pathways responsible, even if associated genes are known. One main reason for this lack of success is the complex genotype-to-phenotype relationships among diseases and their associated genes and mutations. In particular, (i) the same gene can be associated with multiple disorders (gene pleiotropy); and (ii) mutations in any one of many genes can cause the same clinical disorder (locus heterogeneity). For example, mutations in *TP53* are linked to 32 clinically distinguishable forms of cancer and cancer-related disorders, whereas mutations in any of at least 12 different genes can lead to long QT syndrome.

With the publication of several large-scale protein-protein interaction networks in human[5–8], researchers have recently begun to use complex cellular networks to explore these genotype-to-phenotype relationships[2,9], on the basis that many proteins function by interacting with other proteins. However, most analyses model proteins as graph-theoretical nodes, ignoring the structural details of individual proteins and the spatial constraints of their interactions. Here, we investigate on a large-scale the underlying molecular mechanisms for the complex genotype-to-phenotype relationships by integrating three-dimensional (3D) atomic-level protein structure information with high-quality large-scale protein-protein interaction data. Within the framework of this structurally resolved protein interactome, we examine the relationships among human diseases and their associated genes and mutations.

RESULTS

Structurally resolved protein interactome for human disease

We first combined 12,577 reliable literature-curated binary interactions filtered from six widely used databases[10–15] (Online Methods) and 8,173 well-verified, high-throughput, yeast two-hybrid (Y2H) interactions[5–8] to produce the high-quality human protein interaction network (hPIN) with 20,614 interactions between 7,401 proteins (**Fig. 1a**).

Next, we structurally resolved the interfaces of these interactions using a homology modeling approach[16]. We used both iPfam[17] and 3did[18] to identify the interfaces of two interacting proteins by mapping them to known atomic-resolution 3D structures of interactions in the Protein Data Bank (PDB)[19] (**Fig. 1a**). Only those interactions in which the interacting domains of both partners (or their homologs) can be found in a 3D structure of an interaction were kept, resulting in a human structural interaction network (hSIN) of 4,222 structurally resolved interactions between 2,816 proteins (**Fig. 1a**). Here, we carefully selected high-quality direct physical interactions between human proteins because interaction databases often contain low-quality and/or nonbinary interactions[20–22], for which interaction interfaces do not exist.

Finally, to compile a comprehensive list of disease-associated genes and their mutations, we combined information from both Online Mendelian Inheritance in Man (OMIM)[23] and the Human Gene

[1]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, USA. [2]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, New York, USA. [3]Department of Medicine, Weill Cornell College of Medicine, New York, New York, USA. [4]Department of Bioinformatics, Maastricht University, Maastricht, The Netherlands. [5]These authors contributed equally to this work. Correspondence should be addressed to H.Y. (haiyuan.yu@cornell.edu).

Received 11 October 2011; accepted 19 December 2011; published online 15 January 2012; doi:10.1038/nbt.2106

Mutation Database (HGMD)[24] (**Fig. 1a**). In total, we were able to collect 62,663 Mendelian mutations in 3,949 protein-coding genes associated with 3,453 clinically distinct disorders (**Supplementary Note 1**), of which 21,716 mutations in 624 disease-associated genes were mapped to corresponding proteins in hSIN (**Fig. 1a**,**b**). All interaction and disease-association data sets are available on our website: http://www.yulab.org/DiseaseInt/.

To evaluate the reliability of our homology modeling approach, we cross-validated domain-domain interactions in 1,456 interactions with co-crystal structures and found that >90% can be correctly inferred from their homologous domains of other interacting pairs in the data set (**Supplementary Note 2**). To further verify the quality of hPIN and hSIN, we investigated enrichment of highly co-expressed and functionally similar[25] interacting pairs in these networks as well as unfiltered interactions relative to random pairs (**Supplementary Note 3**). We found that hPIN is significantly more enriched for co-expressed and functionally similar pairs than unfiltered interactions ($P = 0.002$ and $P < 10^{-20}$ by cumulative binomial tests, respectively; **Fig. 1c**,**d**), verifying the high quality of hPIN and our filtering process. More importantly, hSIN is even more enriched ($P < 10^{-13}$ and $P < 10^{-20}$ by cumulative binomial tests, respectively; **Fig. 1c**,**d**), illustrating the importance of structural resolution.

Enrichment of in-frame disease mutations on interfaces

Disease mutations can be classified into two broad categories—in-frame mutations (including missense point mutations and in-frame insertions or deletions) and truncating mutations (including nonsense point mutations and frameshift insertions or deletions). Disease alleles with in-frame mutations are likely to produce full-length proteins with local defects, whereas those with truncating mutations will only give rise to incomplete fragments. Our list comprises 12,059 in-frame mutations and 9,657 truncating mutations from 624 genes in hSIN.

Although individual experiments have shown that in-frame mutations can lead to loss of interactions[26], previous studies have

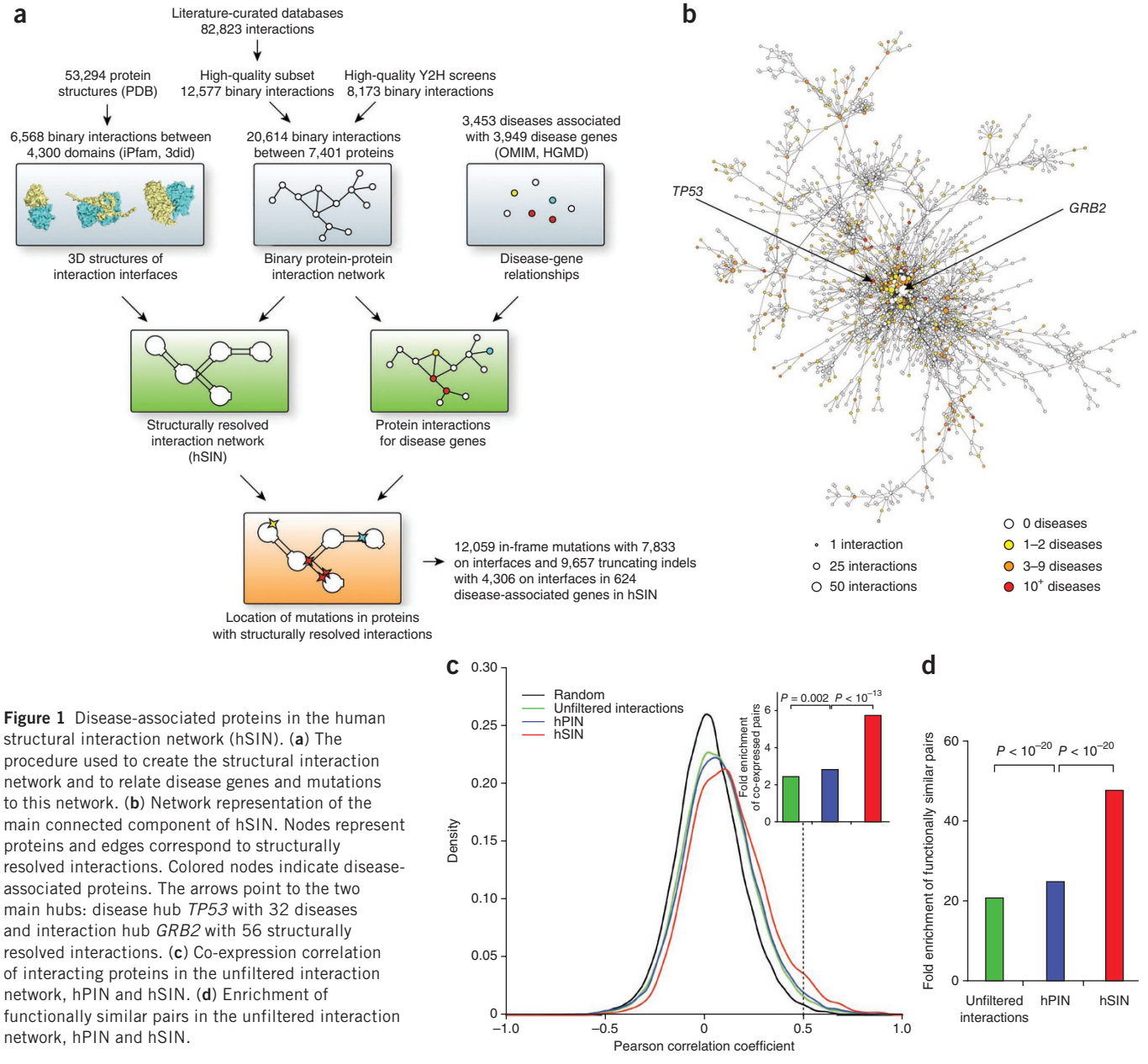

Figure 1 Disease-associated proteins in the human structural interaction network (hSIN). (**a**) The procedure used to create the structural interaction network and to relate disease genes and mutations to this network. (**b**) Network representation of the main connected component of hSIN. Nodes represent proteins and edges correspond to structurally resolved interactions. Colored nodes indicate disease-associated proteins. The arrows point to the two main hubs: disease hub *TP53* with 32 diseases and interaction hub *GRB2* with 56 structurally resolved interactions. (**c**) Co-expression correlation of interacting proteins in the unfiltered interaction network, hPIN and hSIN. (**d**) Enrichment of functionally similar pairs in the unfiltered interaction network, hPIN and hSIN.

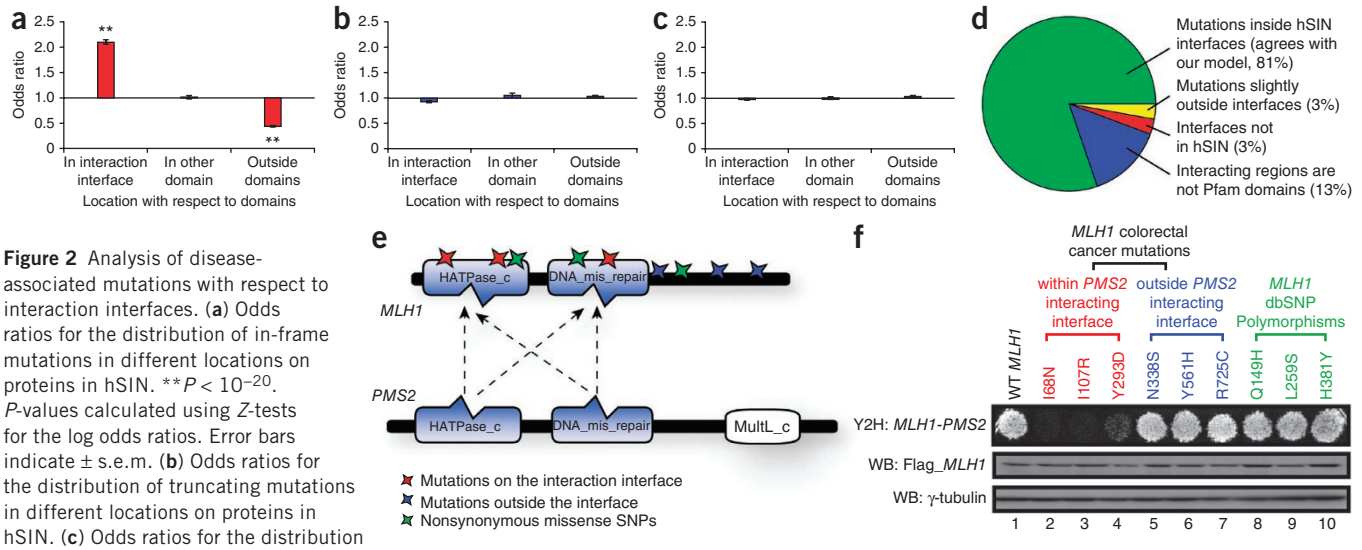

Figure 2 Analysis of disease-associated mutations with respect to interaction interfaces. (**a**) Odds ratios for the distribution of in-frame mutations in different locations on proteins in hSIN. $^{**}P < 10^{-20}$. *P*-values calculated using *Z*-tests for the log odds ratios. Error bars indicate ± s.e.m. (**b**) Odds ratios for the distribution of truncating mutations in different locations on proteins in hSIN. (**c**) Odds ratios for the distribution of nonsynonymous SNPs in different locations on proteins in hSIN. (**d**) Comparison of hSIN with mutations known to modify protein-protein interactions. (**e**) Illustration of *MLH1* and *PMS2* interaction interfaces. Colored stars indicate locations of experimentally tested in-frame mutations and SNPs. (**f**) Effects of in-frame mutations and SNPs on the *MLH1-PMS2* interaction tested by Y2H. Flag-tagged wild-type and mutant *MLH1* were expressed in HEK293T cells, western blot analysis showed similar levels of *MLH1* proteins. γ-tubulin was used as a loading control.

concluded that only a small fraction of disease-associated mutations are expected to specifically affect protein-protein interactions[27,28]. To explore the relationships between mutations and their associated disorders, we investigated positions of the disease-associated mutations with regard to interaction interfaces on the corresponding proteins. Among the 12,059 in-frame mutations, we found that 7,833 are located on interaction interfaces, which is significantly enriched with respect to the relative length of interfaces to whole proteins (odds ratio = 2.1, $P < 10^{-20}$ with a *Z*-test; **Fig. 2a**). In contrast, an enrichment of in-frame mutations was not detected in noninteracting domains (odds ratio = 1.0, $P = 0.70$ with a *Z*-test; **Fig. 2a**). This indicates that specific alteration (disruption or enhancement; **Supplementary Note 4**) of protein-protein interactions plays an important role in the pathogenesis of many disease genes, more than previously expected[27] (**Supplementary Note 5**). On the other hand, truncating mutations seem to be distributed randomly throughout the protein (**Fig. 2b**). We also examined the distribution of 13,783 nonsynonymous single-nucleotide polymorphisms (SNPs)[29] in 806 disease genes in hSIN and found that they, too, are randomly distributed (**Fig. 2c** and **Supplementary Note 6**). These results further confirm our conclusion because alleles with truncating mutations are more likely to produce nonfunctional products[26] and most SNPs in dbSNP are considered to be nondisease-related[30].

To verify that the in-frame mutations on the interfaces in hSIN can interfere with protein interactions, we manually compared them with an independent list of known interaction-altering missense mutations that could be mapped to genes in hSIN[27]. The majority (81%) of these mutations (72 mutations in total) are indeed localized on the interaction interfaces according to hSIN (**Fig. 2d**), confirming the coverage and quality of hSIN (**Supplementary Note 7**).

We also experimentally evaluated the effects of disease-associated mutations and nondisease-related SNPs found in *MLH1*, a well-characterized human DNA mismatch repair gene frequently mutated in hereditary nonpolyposis colorectal cancer[31]. *MLH1* is known to interact with many proteins, including its heterodimeric partner *PMS2*, but the structural basis of most interactions, including with *PMS2*, still remains unknown. Our hSIN predicts that the HATPase_c domain and the DNA_mis_repair domain on *MLH1* are potentially responsible for *MLH1*'s interaction with *PMS2* (**Fig. 2e**). Therefore we hypothesized that mutations within these two domains are likely to alter this interaction. To test our hypothesis, we used Y2H to test six different in-frame, colorectal cancer–associated mutations and three nonsynonymous SNPs found in *MLH1* for their abilities to alter the *MLH1-PMS2* interaction (**Supplementary Note 8** and **Supplementary Fig. 1**). Compared to the wild-type *MLH1*, only missense mutations (I68N, I107R and Y293D) within the predicted *PMS2* interacting interface greatly reduce the *MLH1-PMS2* interaction (**Fig. 2f**). These experimental results further confirm the validity of our predicted interaction interfaces in hSIN. Moreover, they show that in-frame mutations enriched on interfaces could indeed alter corresponding interactions.

Pleiotropy of disease genes—effects of mutations on different interaction interfaces of the same protein

Disease genes are often associated with multiple clinically distinct disorders[2]. To investigate how mutations in the same gene can cause different phenotypes, we examined the relationships between potentially interaction-altering, in-frame, disease-associated mutations within our atomic-resolution structural interaction network, hSIN.

By analyzing the distribution of in-frame mutation pairs on the same gene (**Supplementary Note 9**), we found that in-frame mutation pairs on different interaction interfaces are more than twice as likely to cause different disorders as those on the same interface (46% and 21%, respectively, $P < 10^{-20}$ by a cumulative binomial test; **Fig. 3a**). This suggests that the number of interactions and interfaces are key to understanding the pleiotropic effects of disease genes. Mutations on interaction interfaces of the same protein, mediating different interactions, are more likely to cause distinct interruptions in the overall interactome and can therefore result in different biological consequences and lead to pleiotropic effects. Interestingly, there is no such difference between mutations in different noninteracting domains, further underscoring the importance of protein-protein interactions and their role in understanding disease.

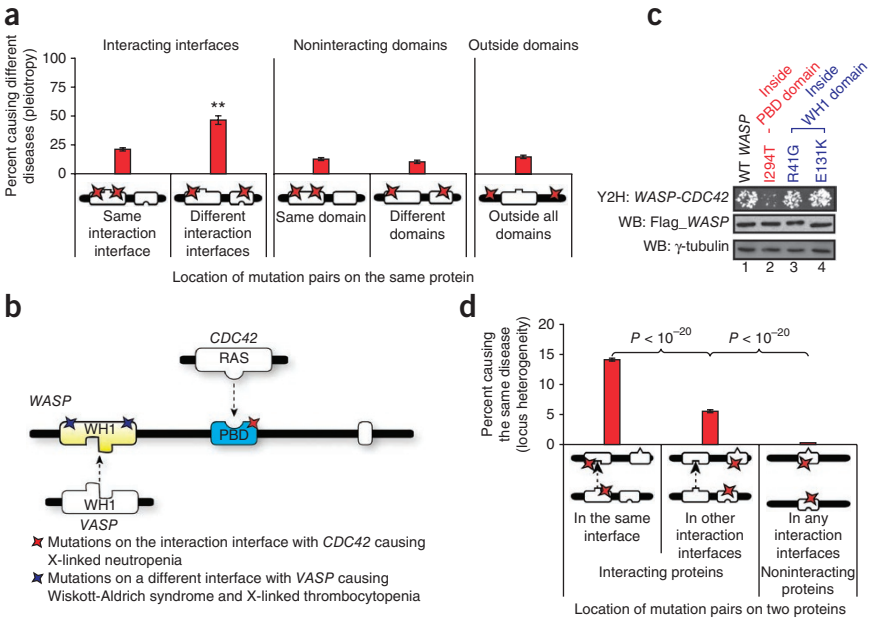

Figure 3 Analysis of pleiotropy and locus heterogeneity. (**a**) Fraction of mutation pairs on the same protein causing different diseases. **, $P < 10^{-20}$. *P*-values calculated using binomial tests. (**b**) Illustration of *WASP* and its interaction interfaces with *CDC42* and *VASP*. Colored stars indicate locations of experimentally tested mutations. (**c**) Effects on the *WASP-CDC42* interaction by mutations on different interaction interfaces tested by Y2H. Flag-tagged wild-type and mutant *WASP* were expressed in HEK293T cells, western blot analysis showed similar levels of *WASP* proteins. γ-tubulin was used as a loading control. (**d**) Fraction of mutation pairs on two proteins causing the same disease.

One well-studied example of pleiotropy is the Wiskott-Aldrich syndrome protein (*WASP*, also known as *WAS*)[32], which contains a WH1 and a PBD domain (**Fig. 3b**). Mutations in this protein can give rise to three diseases: Wiskott-Aldrich syndrome (WAS), X-linked thrombocytopenia (XLT) or X-linked neutropenia (XLN). WAS and XLT are related diseases with XLT being a milder form of WAS, both of which are clinically distinct from XLN (**Supplementary Note 10**). Based on our 3D structural analysis using hSIN, mutations associated with WAS and XLT are in or around the WH1 domain, which is responsible for interaction with *VASP*; mutations for XLN on the other hand are all inside the PBD domain, which performs an entirely different function by interacting with *CDC42* and regulating the autoinhibition and potentially the localization of *WASP*[33–35] (**Fig. 3b**). More interestingly, our experimental results confirm that mutations on different interfaces of *WASP* function differently in terms of altering protein interactions. Specifically, we compared interactions of *CDC42* with the wild-type *WASP* and three disease-associated variants using Y2H. Neither mutation (R41G and E131K; associated with WAS/XLT) located within WH1 domain affects *WASP*'s interaction with *CDC42* (**Fig. 3c**, lanes 3 and 4). However, this is, to our knowledge, the first experimental evidence that one amino acid change within the PBD domain (I294T; associated with XLN) greatly reduces the *WASP-CDC42* interaction (**Fig. 3c**, lane 2). Previous *in vitro* analysis has shown that I294T increases *WASP* activity[36]; our result suggests that I294T might function by disrupting the *WASP-CDC42* interaction, therefore affecting *WASP*'s regulation by *CDC42*.

Locus heterogeneity—effects of mutations on the corresponding interfaces of two interacting proteins

Uncovering the mechanisms through which mutations in different genes can lead to the same disease is critical in finding novel disease-associated genes and ultimately understanding and treating the corresponding disease. Based on the widely accepted 'guilt-by-association' principle, interacting proteins have been shown to have a tendency to share similar functions and cause the same disorders[37]. Earlier implementations of this idea had a significant impact and led to the determination of important disease associations for genes[38]. However, the fraction of successful predictions is still relatively small[39]. One main reason is that most interacting protein pairs share only a subset of their associated disorders.

To understand the underlying molecular mechanism for this phenomenon, we calculated the distribution of in-frame mutation pairs on two different proteins that cause the same disorder (**Supplementary Note 9**). We found, in agreement with previous studies[2], that in-frame mutations on interacting proteins are generally much more likely to cause the same disorder (12%) than random expectation (0.17%, $P < 10^{-20}$ by a cumulative binomial test; **Fig. 3d**). More importantly, our results show that the likelihood for two in-frame mutations on the corresponding interfaces of the interacting proteins to cause the same disorder (14%) is significantly higher than that for two in-frame mutations on two interfaces not mediating their interaction (5.6%, $P < 10^{-20}$ by a cumulative binomial test; **Fig. 3d**). These results further indicate that alteration of specific interactions, caused by mutations on corresponding interfaces of two interacting proteins, plays an important role in the pathogenesis of the same disorder. An interesting example is the hemolytic uremic syndrome, which is associated with mutations on the corresponding interaction interfaces of both *CFH* and *C3* that mediate the interaction between the two proteins[40] (**Supplementary Note 11** and **Supplementary Fig. 2**).

Modeling potential molecular mechanisms of disease genes

Our 3D structural analysis provides potential atomic-level understanding for some of the complex genotype-to-phenotype relationships. More importantly, these results enable us to generate a concrete molecular-mechanism hypothesis for mutations of a certain disorder enriched on a specific interaction interface. For example, they may cause their associated disorders by altering the interactions mediated by the corresponding interfaces (**Fig. 4a**, **Supplementary Fig. 3** and **Supplementary Note 4**). Based on this proposed model, we can further predict new disease-associated genes (that is, those that interact with known disease genes through the interfaces enriched with mutations associated with a certain disease; **Supplementary Note 12** and **Supplementary Fig. 4**). Therefore, our analysis provides a much higher resolution application of the guilt-by-association principle. We then applied this principle to uncover unknown disease-associated genes using hSIN. For each disease, we selected proteins in hSIN that have at least three mutations associated with a certain disease and at least 1.5-fold enrichment on interaction interfaces (Online Methods and **Supplementary Note 13**). Other proteins interacting through the interfaces with enriched disease-specific mutations are predicted to be associated with the corresponding disease. In total,

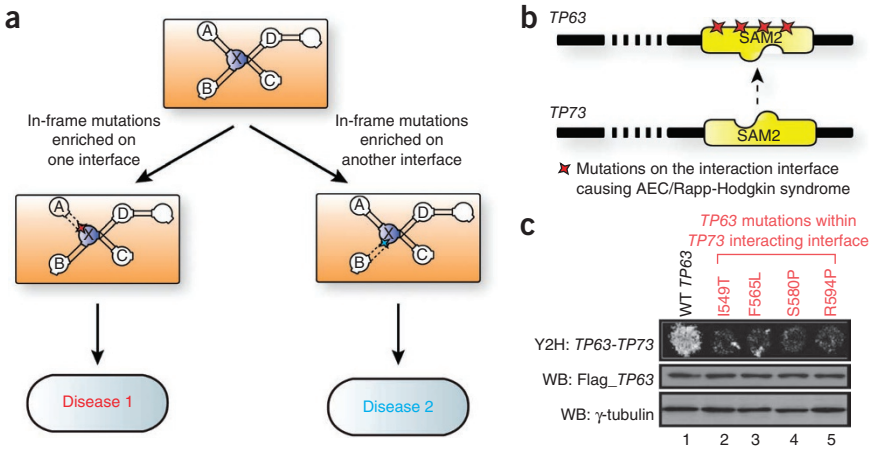

Figure 4 Modeling molecular mechanisms of disease genes and mutations through our structurally resolved interaction network. (**a**) Schematic illustration of using hSIN to understand complex genotype-to-phenotype relationships. In-frame mutations enriched on an interaction interface of protein X likely alter the interaction between protein X and A, leading to one disease, whereas mutations enriched on a different interface are likely to alter the interaction between X and B, leading to another disease. Interactions between protein X and C, as well as X and D are likely to be intact under both scenarios. (**b**) Illustration of *TP63* and its predicted interaction interface with *TP73*. Colored stars indicate locations of experimentally tested mutations. (**c**) Effects on the *TP63-TP73* interaction by mutations on the predicted interacting interface tested by Y2H. Flag-tagged wild-type and mutant *TP63* were expressed in HEK293T cells, western blot analysis showed similar levels of *TP63* proteins. γ-tubulin was used as a loading control.

we predicted 292 new disease genes for 182 different diseases, representing 694 novel disease-to-gene associations. Using threefold cross-validation, we confirmed that our structurally resolved interactome greatly improves the performance of predicting disease-associated genes, compared with existing interaction networks where proteins are modeled as simple graph-theoretical nodes (**Supplementary Note 13** and **Supplementary Figs. 5** and **6**).

To further experimentally validate our predictions, we examined the *TP63-TP73* interaction. *TP63*, unlike its paralog the well-known tumor suppressor gene *TP53*, has an important role in epithelial development[41]. Sequence analysis suggested *TP63* mutations are responsible for Ankyloblepharon-ectodermal defect-cleft lip/palate (AEC) and Rapp-Hodgkin syndrome, two clinically similar disorders (**Supplementary Note 14**)[42]. Interestingly, most of mutations cluster in the SAM2 domain of *TP63*. Based on the known co-crystal structure of *DGKD* homodimer[43], we predict that the SAM2 domain is potentially part of the interface for the *TP63-TP73* interaction (**Fig. 4b**). Therefore, we hypothesized that mutations in the SAM2 domain could affect this interaction. We examined four mutations associated with AEC and/or Rapp-Hodgkin syndrome in the SAM2 domain (I549T, F565L, S580P, R594P) using Y2H. The protein expression levels of the mutants are comparable to the wild-type *TP63* (**Fig. 4c**, middle panel). Our Y2H results indicate that all four mutations substantially reduce the *TP63-TP73* interaction. This suggests that the disruption of proper binding between *TP63* and *TP73* might contribute to the observed phenotypes, and thus *TP73* might also be involved in AEC and/or Rapp-Hodgkin syndrome.

DISCUSSION

From our 3D analysis of disease-associated mutations and their corresponding genes within the atomic-level structurally resolved human protein interactome, we find that specific alteration of protein interactions by in-frame mutations plays an important role in the pathogenesis of many disease genes. More importantly, our results show that the locations of the mutations with respect to the interaction interfaces are crucial in understanding the complex genotype-to-phenotype relationships, including pleiotropy and locus heterogeneity. All observations are demonstrated to be robust to the removal of random interactions and proteins as well as interaction, disease and domain hubs, all of which are potential biases that might be present in our data sets (**Supplementary Note 15** and **Supplementary Figs. 7–21**). Furthermore, all observations remain the same when the calculations are repeated using only known domain-domain interactions from existing co-crystal structures (**Supplementary Note 16** and **Supplementary Fig. 22**).

Our findings are directly applicable to understanding molecular mechanisms of human genetic diseases and discovering new disease-associated genes and mutations both experimentally and computationally, which is of significant interest to both pharmaceutical and medical industries and especially important for treating diseases currently with undruggable target genes. To this end, we provide a list of disease-to-gene associations and generate many hypotheses. Moreover, with the development of exome sequencing, many mutations are being discovered in every study[44]. It is difficult to determine their functional relevance experimentally all at once. Our analysis could potentially provide an approach to prioritize mutations discovered in large-scale sequencing projects, especially for protein pairs without known co-crystal structures.

The construction of our structurally resolved protein interactome largely relies on the availability of 3D co-crystal structures, which limits the coverage of our network. However with the rapid growth of PDB[45], more co-crystal information will become available and the same principles that we developed here can be readily applied to uncover potential molecular mechanisms of many more disease genes whose structural information is currently missing. Another limiting factor is that some interaction interfaces fall outside of the known domain structures, including the disordered regions[46]. Incorporating this type of information will further improve the coverage of hSIN. Moreover, other parts of the protein, especially regions immediately outside of the interacting domains we predicted, might also contribute to the interaction directly or contribute to the correct folding of the corresponding domains. For example, a previous study indicated that the SAM2 domain alone might not be sufficient for the *TP63-TP73* interaction and suggested that residues upstream and downstream of the SAM2 domain and the P53_tetramer domain could also be involved in the interaction[47]. Accordingly, based on the known co-crystal structure of *TP53* homodimer[48], we also predicted in hSIN that the P53_tetramer domain of *TP63* could also be part of the interface for this interaction.

Although we have shown that the interaction pairs in hSIN have significantly higher co-expression correlation and functional similarity in general, further studies can be carried out by considering gene

expression under disease-specific conditions and/or within corresponding tissues for specific disorders. Moreover, study of changes in the protein-protein interaction network during disease progression can also assist the identification of disease biomarkers and modules[49]. In addition to genetic mutations, many other factors including environmental stress, epigenetic modifications and invasion of pathogens might also contribute to human clinical disorders[50]. Integrating these factors in the follow-up studies of the hypotheses generated by our analysis will likely expand our understanding of many human genetic disorders in the near future.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This work was supported by the Startup Fund from Cornell University (to H.Y.), National Cancer Institute R01 CA098626 (to S.M.L.), R21 CA122937 (to S.M.L.) and a generous donation by M. Bell (to S.M.L.). J.D. is supported by the Tata Graduate Fellowship. We thank A. Paccanaro, K. Salehi-Ashtiani, M.A. Yildirim and the anonymous reviewers for critical reading and constructive comments of the manuscript.

AUTHOR CONTRIBUTIONS

H.Y. conceived the study, designed all analyses and oversaw all aspects of the project. X. Wang and B.T. performed all computational analyses, interpreted the results and prepared all figures with J.D.'s help. B.T. and J.D. designed the supplementary website. X. Wei, S.M.L. and H.Y. designed all experiments. X. Wei did all experiments and interpreted the results. H.Y. wrote the manuscript with contributions from all authors. X. Wang and B.T. wrote the supplementary materials. X. Wei wrote all parts related to the background, interpretations and discussion of the experimental results and protocols.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Pasternak, J. *An Introduction to Human Molecular Genetics*, edn. 2 (Wiley, Hoboken, NJ, 2005).
2. Goh, K.I. *et al.* The human disease network. *Proc. Natl. Acad. Sci. USA* **104**, 8685–8690 (2007).
3. Dermitzakis, E.T. & Clark, A.G. Genetics. Life after GWA studies. *Science* **326**, 239–240 (2009).
4. Yildirim, M.A., Goh, K.I., Cusick, M.E., Barabasi, A.L. & Vidal, M. Drug-target network. *Nat. Biotechnol.* **25**, 1119–1126 (2007).
5. Rual, J.F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
6. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
7. Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90 (2009).
8. Yu, H. *et al.* Next-generation sequencing to generate interactome datasets. *Nat. Methods* **8**, 478–480 (2011).
9. Feldman, I., Rzhetsky, A. & Vitkup, D. Network properties of genes harboring inherited disease mutations. *Proc. Natl. Acad. Sci. USA* **105**, 4323–4328 (2008).
10. Keshava Prasad, T.S. *et al.* Human Protein Reference Database–2009 update. *Nucleic Acids Res.* **37**, D767–D772 (2009).
11. Breitkreutz, B.J. *et al.* The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.* **36**, D637–D640 (2008).
12. Aranda, B. *et al.* The IntAct molecular interaction database in 2010. *Nucleic Acids Res.* **38**, D525–D531 (2010).
13. Ceol, A. *et al.* MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.* **38**, D532–D539 (2010).
14. Hu, Z. *et al.* VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Res.* **37**, W115–W121 (2009).
15. Turner, B. *et al.* iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database (Oxford)* **2010**, baq023 (2010).
16. Kim, P.M., Lu, L.J., Xia, Y. & Gerstein, M.B. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314**, 1938–1941 (2006).
17. Finn, R.D., Marshall, M. & Bateman, A. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* **21**, 410–412 (2005).
18. Stein, A., Panjkovich, A. & Aloy, P. 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Res.* **37**, D300–D304 (2009).
19. Berman, H.M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
20. Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110 (2008).
21. Cusick, M.E. *et al.* Literature-curated protein interaction datasets. *Nat. Methods* **6**, 39–46 (2009).
22. Turinsky, A.L., Razick, S., Turner, B., Donaldson, I.M. & Wodak, S.J. Literature curation of protein interactions: measuring agreement across major public databases. *Database (Oxford)* **2010**, baq026 (2010).
23. Amberger, J., Bocchini, C.A., Scott, A.F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.* **37**, D793–D796 (2009).
24. Stenson, P.D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med.* **1**, 13 (2009).
25. Yu, H., Jansen, R., Stolovitzky, G. & Gerstein, M. Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications. *Bioinformatics* **23**, 2163–2173 (2007).
26. Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.* **5**, 321 (2009).
27. Schuster-Bockler, B. & Bateman, A. Protein interactions in human genetic diseases. *Genome Biol.* **9**, R9 (2008).
28. Ferrer-Costa, C., Orozco, M. & de la Cruz, X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J. Mol. Biol.* **315**, 771–786 (2002).
29. Smigielski, E.M., Sirotkin, K., Ward, M. & Sherry, S.T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.* **28**, 352–355 (2000).
30. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
31. Peltomaki, P. & Vasen, H.F. Mutations predisposing to hereditary nonpolyposis colorectal cancer: database and results of a collaborative study. The International Collaborative Group on Hereditary Nonpolyposis Colorectal Cancer. *Gastroenterology* **113**, 1146–1158 (1997).
32. Thrasher, A.J. & Burns, S.O. WASP: a key immunological multitasker. *Nat. Rev. Immunol.* **10**, 182–192 (2010).
33. Kim, A.S., Kakalis, L.T., Abdul-Manan, N., Liu, G.A. & Rosen, M.K. Autoinhibition and activation mechanisms of the Wiskott-Aldrich syndrome protein. *Nature* **404**, 151–158 (2000).
34. Higgs, H.N. & Pollard, T.D. Activation by Cdc42 and PIP(2) of Wiskott-Aldrich syndrome protein (WASp) stimulates actin nucleation by Arp2/3 complex. *J. Cell Biol.* **150**, 1311–1320 (2000).
35. Moulding, D.A. *et al.* Unregulated actin polymerization by WASp causes defects of mitosis and cytokinesis in X-linked neutropenia. *J. Exp. Med.* **204**, 2213–2224 (2007).
36. Ancliff, P.J. *et al.* Two novel activating mutations in the Wiskott-Aldrich syndrome protein result in congenital neutropenia. *Blood* **108**, 2182–2189 (2006).
37. Oliver, S. Guilt-by-association goes global. *Nature* **403**, 601–603 (2000).
38. Wang, X., Gulbahce, N. & Yu, H. Network-based methods for human disease gene prediction. *Brief Funct. Genomics* **10**, 280–293 (2011).
39. Oti, M., Snel, B., Huynen, M.A. & Brunner, H.G. Predicting disease genes using protein-protein interactions. *J. Med. Genet.* **43**, 691–698 (2006).
40. Noris, M. & Remuzzi, G. Atypical hemolytic-uremic syndrome. *N. Engl. J. Med.* **361**, 1676–1687 (2009).
41. Yang, A. *et al.* p63, a p53 homolog at 3q27–29, encodes multiple products with transactivating, death-inducing, and dominant-negative activities. *Mol. Cell* **2**, 305–316 (1998).
42. Bougeard, G., Hadj-Rabia, S., Faivre, L., Sarafan-Vasseur, N. & Frebourg, T. The Rapp-Hodgkin syndrome results from mutations of the TP63 gene. *Eur. J. Hum. Genet.* **11**, 700–704 (2003).
43. Harada, B.T. *et al.* Regulation of enzyme localization by polymerization: polymer formation by the SAM domain of diacylglycerol kinase delta1. *Structure* **16**, 380–387 (2008).
44. Bamshad, M.J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).
45. Chandonia, J.M. & Brenner, S.E. The impact of structural genomics: expectations and outcomes. *Science* **311**, 347–351 (2006).
46. Neduva, V. *et al.* Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol.* **3**, e405 (2005).
47. Chi, S.W., Ayed, A. & Arrowsmith, C.H. Solution structure of a conserved C-terminal domain of p73 with structural homology to the SAM domain. *EMBO J.* **18**, 4438–4445 (1999).
48. Clore, G.M. *et al.* Refined solution structure of the oligomerization domain of the tumour suppressor p53. *Nat. Struct. Biol.* **2**, 321–333 (1995).
49. Hwang, D. *et al.* A systems approach to prion disease. *Mol. Syst. Biol.* **5**, 252 (2009).
50. Vidal, M., Cusick, M.E. & Barabasi, A.L. Interactome networks and human disease. *Cell* **144**, 986–998 (2011).

ONLINE METHODS

Compiling a high-quality comprehensive list of diseases, disease-associated genes and mutations. Two databases that contain relationships between genes and diseases were used: Online Mendelian Inheritance in Man (OMIM)[23] and the Human Gene Mutation Database (HGMD)[24]. Because disease names are not standardized between or even within the databases, we performed extensive informatics operations as well as manual curation to combine the two data sources (**Supplementary Note 1**).

Individual mutations with their flanking sequence were translated into amino acid sequences and aligned to the protein sequence (using SwissProt[51] release 57.6, which corresponds to the sequences used by Pfam[52] release 24). From HGMD, all "disease-causing mutations" and "disease-associated polymorphisms of functional significance" were selected, for a total of 74,048 mutations (including both point mutations and insertion and deletions). Of these, 49,785 corresponded exactly to the SwissProt sequence, and an additional 12,878 matched after correcting the numbering; the rest was discarded. For further analysis, we used only those mutations in genes for which we were able to structurally resolve their interactions (21,716 mutations).

The disease-to-gene associations, the location of disease-associated mutations, as well as the structural interaction network can be explored interactively on our website: http://www.yulab.org/DiseaseInt/. We have also included all of our data sets in **Supplementary Tables 1–11**. We will regularly update our data sets and the website to keep up with the growth of the databases used.

Compiling a high-quality, comprehensive list of binary protein-protein interactions. Protein-protein interactions were obtained from these databases: Human Protein Reference Database (HPRD)[10] release 9; BioGrid[11] release 3.0.66; IntAct[13] downloaded July 27, 2010; Molecular Interaction Database (MINT)[13] version of July 22, 2010; VisANT[14] downloaded July 27, 2010; iRefWeb[15] 3.9. An interaction was considered to be of high quality when it fulfilled two criteria: (i) it has at least two separate publications, and (ii) each of these publications needs to have a binary evidence code; that is, the experiments used for determining the interaction must be in principle capable of determining direct, binary protein-protein interactions. All interactions that did not satisfy these criteria were discarded. Many evidence codes used in the databases in support of binary protein-protein interactions represent experimental assays that cannot distinguish between direct or indirect interactions (such as tandem affinity purification). Some experimental assays do not even in principle study protein-protein interactions (such as electrophoretic mobility shift). We therefore manually considered each evidence code to make sure only experiments that produce direct binary protein-protein interactions were used. The curated list of evidence codes is provided in **Supplementary Table 1**.

We also compiled 8,173 high-quality Y2H interactions from reliable data sets that have all been verified with multiple orthogonal assays[5–8]. The union of the high-quality literature-curated and Y2H interactions is called "human protein-protein interaction network" (hPIN) with 20,614 interactions between 7,401 proteins (**Supplementary Table 2**).

Constructing the human structural interaction network. In order to structurally resolve the protein-protein interactions in hPIN, we used known 3D structures of the two proteins in complex, or their close homologs[16]. For each interaction, we determined whether the two interactors contain a Pfam domain pair that has been seen to interact in at least one protein structure in either 3did[18] or iPfam[17]. The set of Pfam domains on protein A that have corresponding interacting domains on protein B is then considered the interaction interface of protein A for protein B. Pfam release 24, iPfam release 21 and 3did release of August 8, 2010 were used. We used only "Pfam A" domains that are both "significant" and "in-full," as defined by Pfam.

The interactions in hSIN then have two independent lines of support: the interaction is known to be genuine based on experimental evidence, and their interaction is structurally resolved either directly from a protein complex structure or from significant homology to such structures. All interactions are listed in **Supplementary Table 3**. Compared to two previous data sets[27,53], our data set has one major advantage: based on our extensive experience in evaluating the quality of binary interactions[7,20,21] we carefully selected high-quality binary interactions, as opposed to merely collecting all interactions reported in the literature. This is extremely important because literature-curated interactions could contain low quality and/or nonbinary ones[20–22]. When two proteins do not bind each other directly, the concept of interaction interfaces does not apply.

Evaluating network quality. To assess the quality of our constructed networks, we downloaded microarray data conducted on various tissues and cell lines, as well as across multiple cell cycles, to compare co-expression correlation of interacting pairs[54–56]. Expression values were carefully normalized and combined[57–59]. We then calculated pair-wise Pearson Correlation Coefficients. To evaluate the functional similarity between interacting pairs, we downloaded Biological Process (BP), Cellular Component (CC), or Molecular Function (MF) branches of the Gene Ontology (GO)[60] and calculated functional similarity scores between protein pairs[25]. The enrichment of co-expressed and functionally similar interacting pairs in the unfiltered interaction network, hPIN and hSIN were calculated and compared at various cutoffs (**Supplementary Note 3** and **Supplementary Table 5**).

Statistical analysis of mutations and SNPs. In addition to the mutations downloaded from HGMD database as described above, we further obtained missense SNPs (functional class "42") for disease genes in hSIN from dbSNP database build 132 (ref. 29). Similarly to the mutation data set, we only kept the SNPs with reference amino acid being correctly mapped to the Pfam sequences in release 24. We finally analyzed 13,783 missense SNPs in 806 disease genes in hSIN (see **Supplementary Fig. 24** for distribution of SNPs in all genes in hSIN). Enrichment of mutations and SNPs on interaction interfaces was established by comparing the observed number of mutations and SNPs on interfaces to the relative length of the coding sequences forming the interfaces (**Supplementary Note 6**). Pair-wise mutation calculations were done by taking all possible pairs of mutations belonging to the different categories under comparison (**Supplementary Note 9**). Sample sizes involved in all calculations are listed in **Supplementary Table 6**.

Predicting disease-to-gene associations. For each disease, we calculated the enrichment of mutations causing that disease in each domain that is part of an interaction interface. It is likely that interactions mediated by interfaces specifically enriched with in-frame mutations are affected by these mutations. Genes with less than three mutations causing a specific disease, as well as interacting domains with an enrichment of <1.5 were discarded. False-discovery rate at 10% was used to correct for multiple comparisons (**Supplementary Note 13**). This provided a list of 194 genes that together contain 480 interacting domains that are specifically enriched with mutations causing a certain disease. Subsequently, we used hSIN to find all proteins that interact on one of those domains (an example is presented in **Supplementary Note 17** and **Supplementary Fig. 23**). If the interaction partner is already known to be associated with the specific disease under investigation the interaction was classified as 'known', as this would not be a novel disease-to-gene association prediction. For the numbers presented in the main text, the known associations were discarded, keeping only the novel predictions. All interactions involving enrichment of mutations on the interface, including the predicted and known associations are listed in **Supplementary Table 4**. The performance of prediction was evaluated using threefold cross-validation (**Supplementary Note 13** and **Supplementary Figs. 5** and **6**).

Construction of plasmids and disease mutant clones. Wild-type *MLH1*, *WASP*, *CDC42*, *TP63* and *TP73* entry clones are from hORFeome 3.1 collection[61]. Wild-type *PMS2* cDNA was purchased from Open Biosystems (clone ID 7939766). To generate disease mutant clones, PCR mutagenesis was done as previously described[26,62]. Briefly, wild-type genes in AD or DB vector were used as templates in PCR reactions to generate N- and C-terminal fragments both containing the desired mutation in their overlapping regions. BP recombination reactions were done according to manufacture's manual (Gateway BP Clonase II enzyme mix) to clone mutant clones into the entry vector (pDONR223). Wild-type *MLH1*, *WASP*, *TP63* and mutant clones were also PCR cloned into the mammalian expression vector pcDNA3 (Invitrogen Life Technologies) using XbaI and NotI restriction sites. Flag-tag was introduced

into the C-terminal end of genes. Mutagenesis and cloning primers used in this study are listed in **Supplementary Table 7**.

Y2H. Y2H was done as previously described[20]. *PMS2*, *CDC42* and *TP73* were transferred into AD vector using Gateway LR reactions. Wild-type/mutant *MLH1*, *WASP* and *TP63* were transferred into DB vector. AD and DB constructs were transformed into Y2H strains *MAT*a Y8800 and *MAT*α Y8930, respectively. Transformed yeast was spotted onto YPD plates and incubated at 30 °C for ~20 h before replica plating onto SC-Leu-Trp plates. These plates were kept at 30 °C for 24 h, then replica plated onto each of the four plates (SC-Leu-Trp-His, SC-Leu-His+CYH, SC-Leu-Trp-Ade, SC-Leu-Ade+CYH), 3 d after plates were scored for protein interactions.

Cell culture and transient expression. HEK293T cells were maintained in complete DMEM supplemented with 10% FBS. HEK293T cells were transfected with Lipofectamine 2000 reagent (Invitrogen) at a 6:1 (liter/gram) ratio with DNA. Cells were harvested 24 h after transfection.

Immunoblotting. Transfected cells were gently washed three times in PBS and then resuspended using 200 µl 1% NP-40 lysis buffer (1% Nonidet P-40, 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 EDTA-free Complete Protease Inhibitor tablet (Roche), 1 M sodium orthovanadate, 1 mM sodium fluoride) per well of 6-well plate for 20 min on ice in Eppendorf tubes. Extracts were cleared by centrifugation for 10 min at 13,000 r.p.m. (>16,000*g*) at 4 °C. Protein lysate (20 µl) were subjected to SDS-PAGE and protein blotting. Anti-Flag (Sigma-Aldrich), anti-γ-tubulin (Sigma-Aldrich T5192) antibodies were used for immunoblotting analyses. Horseradish peroxidase–linked secondary antibodies were purchased from GE Healthcare. Full-length blots are available in **Supplementary Figure 25**.

51. The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **35**, D193–D197 (2007).
52. Finn, R.D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **38**, D211–D222 (2010).
53. Prieto, C. & De Las Rivas, J. Structural domain-domain interactions: assessment and comparison with protein-protein interaction data to improve the interactome. *Proteins* **78**, 109–117 (2010).
54. Whitfield, M.L. *et al.* Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13**, 1977–2000 (2002).
55. Su, A.I. *et al.* Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. USA* **99**, 4465–4470 (2002).
56. Bar-Joseph, Z. *et al.* Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. *Proc. Natl. Acad. Sci. USA* **105**, 955–960 (2008).
57. Irizarry, R.A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
58. Smyth, G.K. & Speed, T. Normalization of cDNA microarray data. *Methods* **31**, 265–273 (2003).
59. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
60. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
61. Lamesch, P. *et al.* hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. *Genomics* **89**, 307–315 (2007).
62. Suzuki, Y. *et al.* A novel high-throughput (HTP) cloning strategy for site-directed designed chimeragenesis and mutation using the Gateway cloning system. *Nucleic Acids Res.* **33**, e109 (2005).

doi:10.1038/nbt.2106