RESEARCH ARTICLE

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

Elucidating Common Structural Features of Human Pathogenic Variations Using Large-Scale Atomic-Resolution Protein Networks

Jishnu Das,[1,2] Hao Ran Lee,[1,2†] Adithya Sagar,[2,3†] Robert Fragoza,[2,4†] Jin Liang,[2] Xiaomu Wei,[2] Xiujuan Wang,[1,2] Matthew Mort,[5] Peter D. Stenson,[5] David N. Cooper,[5] and Haiyuan Yu[1,2*]

[1]*Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853;* [2]*Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, New York 14853;* [3]*Department of Biomedical Engineering, Cornell University, Ithaca, New York 14853;* [4]*Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853;* [5]*Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff CF14 4XN, UK*

Communicated by Mauno Vihinen

Received 26 April 2013; accepted revised manuscript 14 February 2014.

Published online 5 March 2014 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22534

ABSTRACT: With the rapid growth of structural genomics, numerous protein crystal structures have become available. However, the parallel increase in knowledge of the functional principles underlying biological processes, and more specifically the underlying molecular mechanisms of disease, has been less dramatic. This notwithstanding, the study of complex cellular networks has made possible the inference of protein functions on a large scale. Here, we combine the scale of network systems biology with the resolution of traditional structural biology to generate a large-scale atomic-resolution interactome-network comprising 3,398 interactions between 2,890 proteins with a well-defined interaction interface and interface residues for each interaction. Within the framework of this atomic-resolution network, we have explored the structural principles underlying variations causing human-inherited disease. We find that in-frame pathogenic variations are enriched at both the interface and in the interacting domain, suggesting that variations not only at interface "hot-spots," but in the entire interacting domain can result in alterations of interactions. Further, the sites of pathogenic variations are closely related to the biophysical strength of the interactions they perturb. Finally, we show that biochemical alterations consequent to these variations are considerably more disruptive than evolutionary changes, with the most significant alterations at the protein interaction interface.

Hum Mutat 35:585–593, 2014. © 2014 Wiley Periodicals, Inc.

KEY WORDS: network systems biology; structural biology; pathogenic variations; protein–protein interactions

Introduction

The functions of a protein are inherently bound up with its three-dimensional structure—both regular secondary structures and disordered elements play a role in modulating function [Lahiry et al., 2010]. Protein structures are often so intricate that even comparatively minor structural alterations can cause dramatic changes in function. Since such disruptions often lead to disease [Celli et al., 1999; Haberle et al., 2011], a significant amount of effort has been invested in attempting to determine the principles underlying complex structure–function relationships in human proteins. To date, however, most of this effort has been directed toward understanding how individual folds, domains, or structural motifs carry out specific cellular functions [Pearl et al., 2005; Andreeva et al., 2008]. Furthermore, most proteins carry out their functions by interacting with other proteins, all of which are part of a complex cellular network termed the "interactome" [Vidal, 2005; Vidal et al., 2011].

Recently, studies have become focused on how protein networks can be used to infer function and how changes in these networks can lead to human disease [Barabasi et al., 2011; Vidal et al., 2011]. However, these efforts have had only limited success because protein networks are still incomplete [Vidal et al., 2011] and studies to date have treated proteins as mere graph-theoretical points in a mathematical network rather than as biological entities with their own structural details and chemical properties [de Souza, 2012; Wang et al., 2012]. The importance of structural considerations has been well recognized in predicting protein–protein interactions [Tuncbag et al., 2011; Zhang et al., 2012] and functional residues for each interaction [Marks et al., 2012]. However, although structure has been widely employed to understand the evolutionary impact of single-nucleotide polymorphisms (SNPs) [Sunyaev et al., 2001; Bao and Cui, 2005; David et al., 2012], the number of studies that have examined pathogenic variations in a structural context has been limited [Studer et al., 2013]. To address this deficiency, we previously used a domain-level interaction network to show that in-frame pathogenic variations tend to be enriched within interacting domains [Wang et al., 2012]. However, interacting domains comprise not only interface residues that are directly involved in the physical interaction between the two proteins but also other noncontact residues. In our earlier study, we did not differentiate between these two categories of amino acid residues. Since it is generally considered that interface residues mediate protein–protein interactions [Jones and Thornton, 1996], it is of paramount importance to examine the

Additional Supporting Information may be found in the online version of this article.

†These authors contributed equally.

*Correspondence to: Haiyuan Yu, 335 Weill Hall, 237 Tower Road, Ithaca, NY 14853, USA. E-mail: haiyuan.yu@cornell.edu

Contract grant sponsors: National Cancer Institute (grant CA167824); National Institute of General Medical Sciences (grant GM104424); Weill Cornell Medical College Clinical and Translational Science Center Pilot Award; BIOBASE GmbH (through a License Agreement with Cardiff University).

differential distribution of pathogenic variations between interface and noninterface residues within interacting domains. Moreover, only at the resolution of individual amino acid residues is it possible to ascertain structural (i.e., biophysical and biochemical) principles governing pathogenic processes.

To this end, we present here a large-scale atomic-resolution human interactome network by systematically identifying the interaction interfaces and corresponding residues mediating all interactions with available cocrystal structures in the Protein Data Bank (PDB) [Berman et al., 2000]. Using this atomic-resolution interactome network, we analyze the distribution of pathogenic variations in different regions of human proteins focusing on interface and noncontact residues within interacting domains. We also explore how the locational specificity of these variations is directly associated with the strength of the interactions they disrupt. Finally, we examine biochemical properties of human pathogenic variations and compare them with their evolutionary counterparts.

Methods

Calculating Atomic-Resolution Interface Residues for Human Protein Interactions

To calculate atomic-resolution interaction interfaces (Supp. Fig. S1), we systematically examined a comprehensive list of 7,340 PDB cocrystal structures and were able to determine atomic-resolution interaction interfaces for 3,398 unique human protein–protein interactions between 2,890 proteins. To define the interface, we used a water molecule of diameter 1.4 Å as a probe and calculated the relative solvent accessible surface areas of the interacting pair as well as the individual proteins involved in the interaction [Hubbard and Thornton, 1993]. All calculations were performed using Naccess [Hubbard and Thornton, 1993]. Residues whose relative accessibilities changed by more than 1 $Å^2$ were considered as potential interface residues. Amino acids at the interface reside on the surfaces of the corresponding proteins, but tend to become buried in the cocrystal structure as the two proteins bind to each other. It follows that these residues should experience a significant decrease in accessible surface area when the bound and the unbound states of the protein chains are compared [Franzosa and Xia, 2011]. In most cases, our calculations incorporated multiple instances of the same interaction from different chains within the same PDB structure or entirely different PDB structures representing the same interaction. This allows us to accurately determine the exact interface, and normalize differences due to specific crystallization conditions [Chayen and Saridakis, 2008]. We take the union over all such instances subject to the constraint that the particular protein pair contains at least five interface residues for both interacting proteins. This ensures that all the interfaces included in our calculations represent significant regions of molecular contact, eliminating potential crystal contacts. Furthermore, 1,689/3,398 (49.7%) interactions used in this study have been detected by at least one other assay and were reported independently in a separate publication. This confirms that interactions used in this study are not only real but also reproducible using other assays.

To further refine the set of identified interface residues, we required that they be necessarily present on the surface of the protein. To determine which residues were on the surface, we calculated the fraction of surface area for each residue in the individual protein chains that was accessible to the water molecule probe defined above [Hubbard and Thornton, 1993]. If more than 15% of the total surface area for a particular residue was accessible to the water molecule probe, we defined that particular amino acid to be on the surface, otherwise it was considered to be buried. Using these two criteria, for each interaction we obtained a set of 141,686 residues that represent the interface for 3,398 interactions from 7,340 atomic-resolution cocrystal structures. The fraction of homomeric interactions to heteromeric interactions is ~2:1 as the PDB is enriched for homodimers as compared with heterodimers.

Identifying Interacting Domains for Each Interaction

We generated a list of putative interacting domains utilizing the "homology modeling approach" as described earlier [Meyer et al., 2013] using both 3did [Stein et al., 2011] and iPfam [Finn et al., 2005]. However, some of the domain pairs identified as interacting by 3did and iPfam for a particular protein pair may not have been supported by the corresponding cocrystal structure as they may have been inferred from other cocrystal structures. Therefore, to avoid potential false positives, we additionally required that these domains should contain at least one interface residue for them to be considered as interacting domains. Moreover, the set of interacting domains inferred by 3did and iPfam were not always complete. For our analysis, we took advantage of our own atomic-resolution interface calculations to identify a comprehensive set of interacting domains for each cocrystal structure, and included interacting domains not identified by 3did or iPfam if they had five or more interface residues.

Compiling a Comprehensive List of Pathogenic Variations and SNPs

We compiled a comprehensive list of 94,476 pathogenic variations from HGMD [Wang et al., 2012; Stenson et al., 2014] as described earlier [Wang et al., 2012]. We updated our earlier lists with a newer version of the HGMD dataset (HGMD Professional v.2012.4). Specifically, we used in-frame variations (both missense variations and in-frame microinsertions and microdeletions) classified as "DM" in HGMD. For further analysis, we employed a total of 17,306 variations in 673 genes for which we were able to define at least one atomic-resolution interaction interface. We also compiled a set of nonsynonymous SNPs from the Exome Sequencing Project [Fu et al., 2013] from which we derived a dataset of 94,084 SNPs in 2,829 genes for which we were able to define at least one atomic-resolution interaction interface.

Using our publicly available supplementary website, http://www.yulab.org/Supp/AtomInt, researchers can query interface residues for their favorite interaction.

Criteria Used to Choose *PTS–PTS* Homodimeric Interaction for Experimental Validation

The following criteria were used to choose the *PTS–PTS* homodimeric interaction for experimental validation of the effects of pathogenic variants within and outside the interacting domain:

(a) the interaction is supported by a cocrystal structure.
(b) the wild-type *PTS* clone is available in our library.
(c) the wild-type interaction (*PTS–PTS*) is amenable to testing in our yeast two-hybrid (Y2H) system.
(d) there is a pathogenic variation in the interacting domain but outside interface residues.
(e) there is a different pathogenic variation outside both the interface residues and the interacting domain.

Generation of PTS Variants

Wild-type *PTS* is obtained from the human ORFeome v8.1 collection [Yang et al., 2011]. To generate the alleles R25Q and R9C corresponding to two different pathogenic variations, sequence-verified single-colony wild-type *PTS* and corresponding mutagenic primers (designed according to the protocol accompanying the Stratagene QuikChange Site-Directed Mutagenesis Kit #200518) were aliquoted together. Mutagenesis PCR was then performed as specified by the New England Biolabs (NEB) PCR protocol for Phusion polymerase (M0530L), noting that PCR was limited to 18 cycles. Samples were then digested by *Dpn*I (NEB R0176L) according to the manufacturer's manual. After digestion, samples were transformed into competent *Escherichia coli* and then individually streaked onto LB plates containing spectinomycin to obtain single colonies. The generated clones were verified by Sanger sequencing.

Y2H

Y2H was done as previously described [Wang et al., 2012]. Wild-type *PTS* and both pathogenic variant alleles were transferred by Gateway LR reactions into our Y2H pDEST-AD and pDEST-DB vectors. DB-X and AD-Y plasmids were transformed individually into the Y2H strains *MAT*α Y8930 and *MAT*a Y8800, respectively. Each of the DB-X *MAT*α transformants (wild type and variants) were then mated against corresponding AD-Y *MAT*a transformants (wild type and variants), including inoculation of AD-Y and DB-X yeast cultures, mating on YEPD media (incubated overnight at 30°C), and replica plating onto selective Synthetic Complete media lacking leucine, tryptophan, and histidine, and supplemented with 1 mM of 3-amino-1,2,4-triazole (SC-Leu-Trp-His+3AT), SC-Leu-His+3AT plates containing 1 mg/l cycloheximide (SC-Leu-His+3AT+CHX), SC-Leu-Trp-Adenine (Ade) plates, and SC-Leu-Ade+CHX plates to test for CHX-sensitive expression of the *LYS2::GAL1-HIS3* and *GAL2-ADE2* reporter genes. The plates were incubated overnight at 30°C and replica-cleaned the following day. Plates were then incubated for another 3 days, after which positive colonies were scored as those that grow on SC-Leu-Trp-His+3AT and/or on SC-Leu-Trp-Ade, but not on SC-Leu-His+3AT+CHX or on SC-Leu-Ade+CHX. Disruption of an interaction by a variation was defined as significant reduction of growth when compared with the Y2H phenotype of the wild-type *PTS*–*PTS* interaction.

Western Blotting

Wild-type and both *PTS* variants were cloned into MSCV-N-FLAG-HA-IRES-Puro vector [Behrends et al., 2010] and transfected into HEK293T cells to express HA-tagged wild-type and mutated proteins. HEK293T cells were maintained in complete DMEM medium supplemented with 10% fetal bovine serum. Cells were transfected with Lipofectamine 2000 (Invitrogen, Carlsbad, CA) at a 5:1 (μl/μg) ratio with DNA and harvested 24 hr after transfection. Cells were gently washed three times in PBS and then resuspended using 200 μl 1% NP-40 lysis buffer (1% Nonidet P-40, 50 mM Tris–HCl pH 7.5, 150 mM NaCl, 1× EDTA-free Complete Protease Inhibitor tablet [Roche, Indianapolis, IN 05056489001]) and kept on ice for 30 min. Extracts were cleared by centrifugation for 10 min at 15,870 *g* at 4°C. Extracts (25 μl) were mixed with 6× loading buffer and subjected to SDS-PAGE. Proteins were then transferred from the gel onto PVDF membranes (GE Healthcare, Piscataway, NJ RPN303F). Anti-HA (Sigma, St. Louis, MO H9658) and anti-γ-tubulin (Sigma T5192) were used at 1:3,000 dilutions for immunoblotting analysis. Blotting signal was developed with Novex ECL HRP chemiluminescent substrate reagent kit (Invitrogen WP20005) and captured with Amersham Hyperfilm MP (GE Healthcare 28906843).

Results and Discussion

Atomic-Resolution Structural Analysis of Pathogenic Variations and Their Molecular Mechanisms

Pathogenic variations belong to two broad categories—in-frame variations (both missense variations and in-frame microinsertions and microdeletions) and truncating variations (both nonsense point variations and frameshift insertions or deletions) [Zhong et al., 2009]. We previously found that in-frame pathogenic variations are nonrandomly distributed in proteins—indeed, they tend to be enriched within interacting domains. On the other hand, truncating variations do not show any particular trend with regard to their distribution in different parts of the protein [Wang et al., 2012].

It has been commonly accepted that interface residues mediate interactions between proteins [Jones and Thornton, 1996; Hu et al., 2000]. Moreover, it is generally believed that "only a small portion of interface residues, the so-called hot-spot residues, contribute the most to the binding energy of the protein complex" [Assi et al., 2010]. These hot-spots are often the targets of drug molecules [Wells and McClendon, 2007]. Owing to the limits of the resolution of our previous study [Wang et al., 2012], we were able to perform the investigation only at the domain level, not at the level of individual residues. Employing the newly derived atomic-resolution interactome network, we set out to systematically examine whether pathogenic variations tend to specifically alter interface residues, as our previous results suggested that it might be the case. This network is higher resolution that other structurally resolved networks [Wang et al., 2012; Khurana et al., 2013] as it reports not just interacting domains for 3,398 interactions, but individual amino acids residues mediating each interaction. We calculated the enrichment of in-frame variations at the interaction interface, the remainder of the interacting domain, and the rest of the protein. We found that in-frame variations are enriched significantly both at the interface and in the remainder of the interacting domain (odds ratio = 1.67, $P < 10^{-3}$ for interface residues; odds ratio = 1.75, $P < 10^{-3}$ for the remainder of the interacting domain; Fig. 1A, Supp. Note S1). To confirm that the observed trends are robust, we performed the same calculations with only the fraction of the protein in the actual cocrystal, because in many cases the crystallized structure does not contain full-length proteins. 62.6% of all the pathogenic variations used for our calculations in Figure 2A are present within cocrystal structures. Using only these variations, our results remain unchanged—in-frame variations are enriched at both the interface and in the remainder of the interacting domain even if we consider only residues depicted within the cocrystal structures (Supp. Fig. S2A). To assess the significance of a decrease in solvent accessibility, we used randomly chosen cutoffs—decreases of 0.5, 2, and 5 Å^2 in solvent-accessible surface area to define three alternate sets of interface residues. Using these three sets of residues, we repeated our calculations in Figure 1A. We find that our results remain unchanged with all three alternate sets of residues (Supp. Fig. S3). This shows that our results are robust to the choice of cutoff for decrease in solvent-accessible area to define interface residues. In fact, the sets of interface residues are very similar for any cutoff between 0.5 to 5 Å^2.

Our result shows that it is not simply the interface residues, but rather the interacting domain in its entirety that plays an important

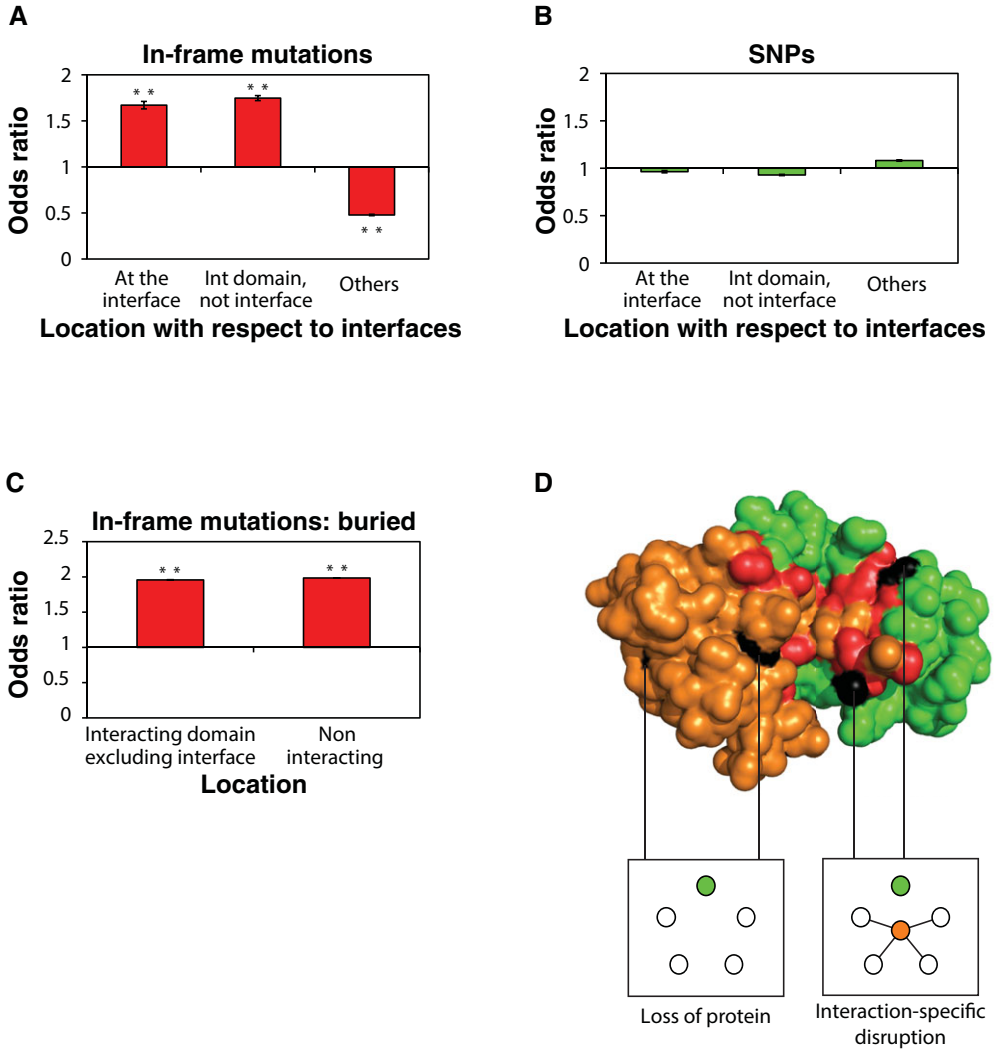


Figure 1. Atomic-resolution structural analysis of pathogenic variations. **A**: Odds ratio for the distribution of in-frame variations in different locations on proteins in our atomic-resolution interactome network. $^{**}P < 10^{-3}$. *P* values calculated using *Z*-tests for the log odds ratio. **B**: Odds ratio for the distribution of nonsynonymous SNPs in different locations on proteins in our atomic-resolution interactome network. (C) Enrichment of in-frame variations in buried residues. $^{**}P < 10^{-3}$. Error bars indicate ±SE. **D**: Different mechanistic modes of disruption for variations in different structural environments—variations at the surface are likely to cause interaction-specific disruptions, whereas those buried in the core of the protein are likely to destabilize the entire protein.

role in pathology for many disease genes. As a negative control, we calculated the distribution of 94,084 missense nonsynonymous SNPs from ESP6500 in 2,829 genes and found that these were distributed randomly across the protein (Fig. 1B). Most genes contain relatively few pathogenic variations and SNPs (Supp. Figs. S4A and S4B). Moreover, there is no significant difference ($P = 0.33$) in the distribution of pathogenic variations and SNPs across various genes (Supp. Figs. S4A and S4B), confirming that the differences observed in the distribution of disease-associated variants and SNPs are not due to gene-specific distribution biases. To further confirm that SNPs are indeed randomly distributed across proteins, we repeated our calculations with only those genes that contain at least one disease-associated variant (i.e., those genes used for the calculations in Fig. 1A) and found that SNPs in these genes are also randomly distributed across the length of the protein (Supp. Fig. S4C). Moreover, even if we consider SNPs present only within cocrystal structures, we find that they are still randomly distributed across proteins (Supp. Fig. S2B).

We also note that in-frame variations outside the interface were enriched in buried residues (Fig. 1C). The importance of buried residues in maintaining the overall stability of the protein is well established [Gromiha et al., 1999]. It has been suggested that in-frame and truncating variations have distinct disruption modes—the former is likely to disrupt specific interactions, whereas the latter usually leads to degradation of the entire protein leading to a loss of all interactors [Zhong et al., 2009]. Our results suggest that even for in-frame variations, the possible molecular mechanisms by which variations at or near the interface (and distant from it) affect protein–protein interactions are likely to be distinct: those at the interface are more likely to alter specific interactions, thereby causing the mutated protein either to lose or acquire specific functions; by contrast, in-frame variations in other noninteracting regions are more likely to disrupt the core of the protein and lead to incorrect folding and/or degradation of the protein, resulting in the loss of all interactions for the mutated protein (Fig. 1D).

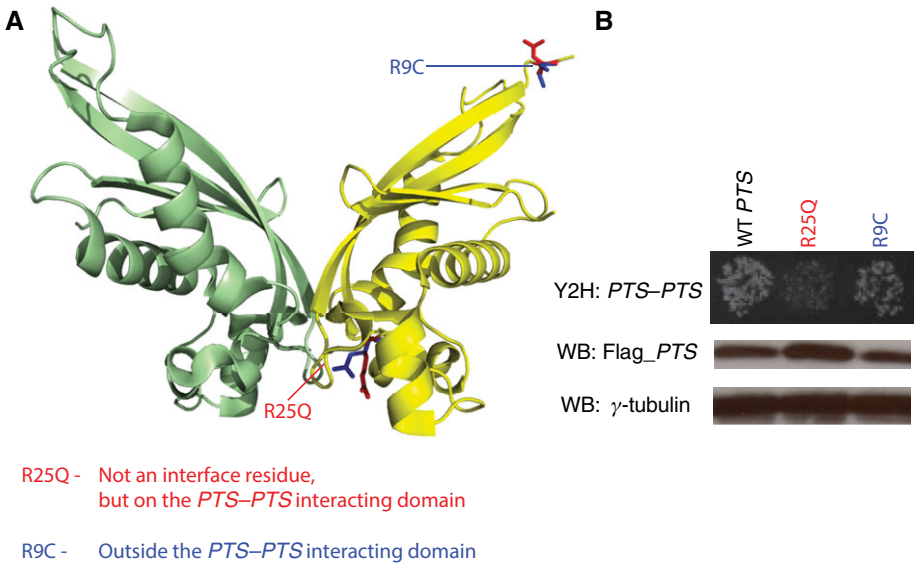


Figure 2. **A**: Crystal structure (PDB id: 3I2B) depicting a R25Q variation in the *PTS–PTS* interacting domain but not at an interface residue and a R9C variation outside the interaction interface. **B**: Y2H assay illustrating that the R25Q variation disrupts the *PTS–PTS* interaction, whereas the R9C variation does not affect the interaction.

To further understand the effects of variations in the interacting domain outside the interface, we examined the effects of two disease-associated variants on the *PTS–PTS* interaction (Fig. 2A). Using site-directed mutagenesis PCR, we introduced the two variants—R25Q and R9C on *PTS*. Although the R25Q variant is located on the PTPS domain that mediates the *PTS–PTS* interaction, it is not at an interface residue. Using Y2H, we confirmed that wild-type *PTS* interacts with itself (Fig. 2B). However, the R25Q variant disrupts this interaction (Fig. 2B). On the other hand, the R9C variant lies outside the interface mediating the *PTS–PTS* interaction. UsingY2H, we confirmed that this variant (R9C) does not affect the interaction (Fig. 2B). This shows that variations in the interacting domain outside the interface can disrupt protein interactions, whereas the same interactions can remain unaffected by variants outside the corresponding interacting domains.

Moreover, using Western blotting, we confirm that all three variants are stable (Fig. 2B). Together, these results show that the R25Q variant causes an interaction-specific disruption—the *PTS–PTS* homodimeric interaction is lost due to a local structural alteration in the corresponding interacting domain. It has been previously shown that the enzymatic activity of the R25Q variant of *PTS* is reduced, but not completely abolished compared with the activity of wild-type *PTS* [Thony et al., 1994; Oppliger et al., 1995]. Our results suggest a molecular mechanistic basis for this reduction—since the dimerization of *PTS* is important for its enzymatic activity [Thony et al., 1994; Oppliger et al., 1995], the pathogenic R25Q that disrupts the *PTS* homodimer reduces this activity. However, since the variant is stable, *PTS* still maintains part of its activity.

Pathogenic Variation Loci Associated with Interaction Strength

To understand the biophysical mechanisms by which in-frame pathogenic variations alter specific interactions, we examined the relationship between the spatial distribution of the variations and the strength of the interactions they perturb. Here, we explored the biophysical strength of an interaction—the stronger the interaction, the higher the free energy difference between the bound and unbound states of the proteins [Noskov and Lim, 2001; Shi et al., 2006]—by calculating the buried surface area of all the interactions in the atomic-resolution human interactome network (Supp. Table S1). The most direct measure of interaction strength is the equilibrium association constant (K_a, inverse of the equilibrium dissociation constant K_d). However, it is difficult to measure K_a in a high-throughput fashion and the amount of experimental K_a data is limited to a handful of human protein–protein interactions.

It has been suggested that the strength of an interaction can be measured by its buried surface area in the cocrystal structure [Jones and Thornton, 1996]. To validate this postulate, we classified all interactions in the network into three distinct categories on the basis of their buried surface area—low, medium, and high (Supp. Note S2). Using a genome-wide microarray analysis that measures the expression levels of human genes at different time points in the cell cycle [Whitfield et al., 2002], we calculated the enrichment in coexpression of proteins involved in these interactions. We found that interactions with high buried surface area are significantly more likely to be coexpressed than interactions with low buried surface area ($P = 0.015$, Fig. 3A, and Supp. Note S3). It is well known that strong, stable interactions are more likely to be coexpressed than weak, transient interactions [von Mering et al., 2002; Yu et al., 2008]. Our result confirms that protein–protein interactions mediated by high buried surface area are indeed stronger. Moreover, we calculated the fraction of these binary interactions independently for the three categories detected in stable protein complexes (Supp. Note S3). We found that interactions with high buried surface area are significantly enriched in stable complexes, further supporting the conclusion that these are stronger interactions (Fig. 3B). Finally, we calculated the correlation between K_a and buried surface area using SKEMPI, a database of binding free energy changes for interactions with supporting cocrystal structures [Moal and Fernandez-Recio, 2012]. For all interactions in SKEMPI involving wild-type human proteins, we calculated the correlation between K_a values and the buried surface area. We find that there is a significant correlation

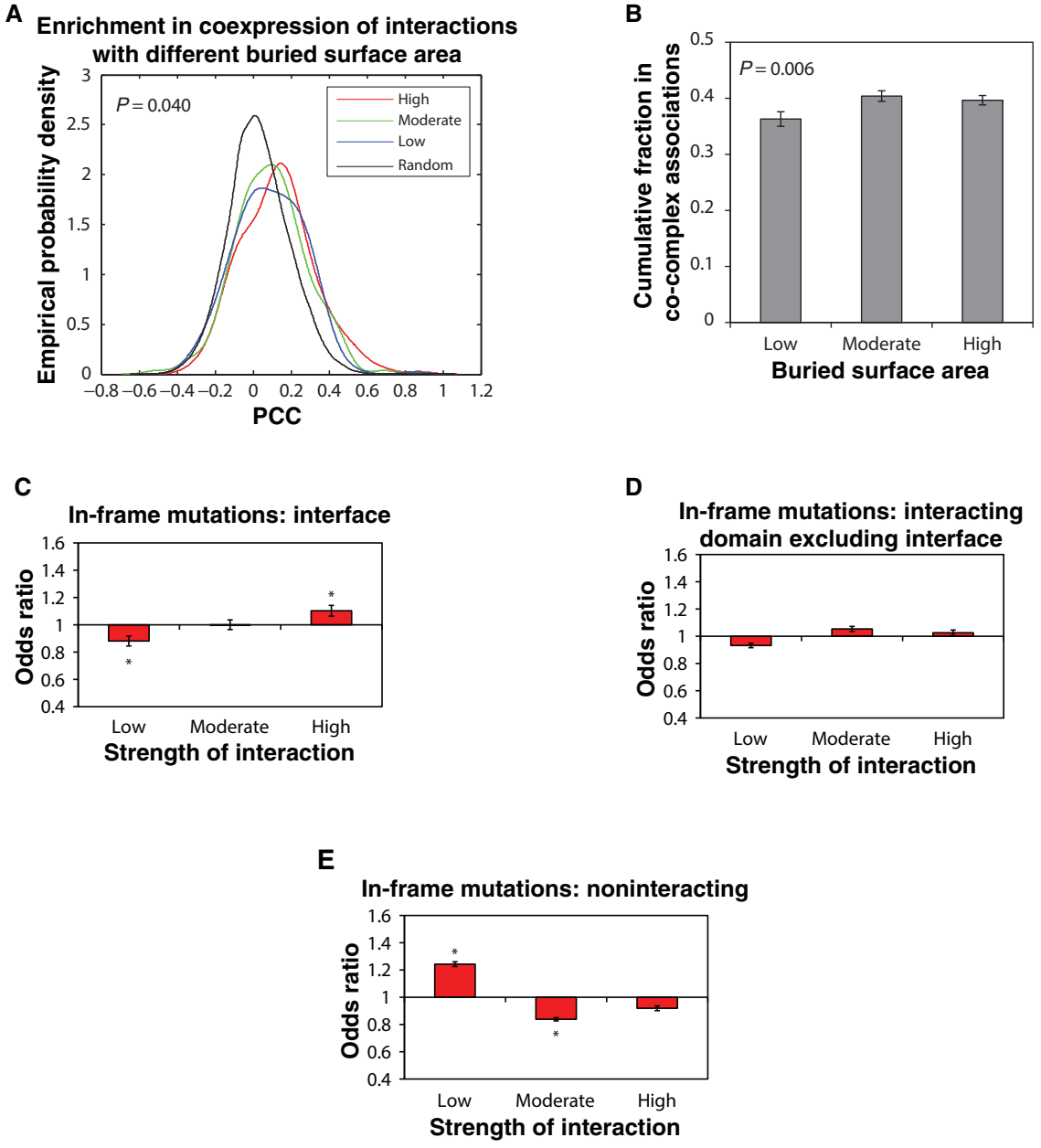
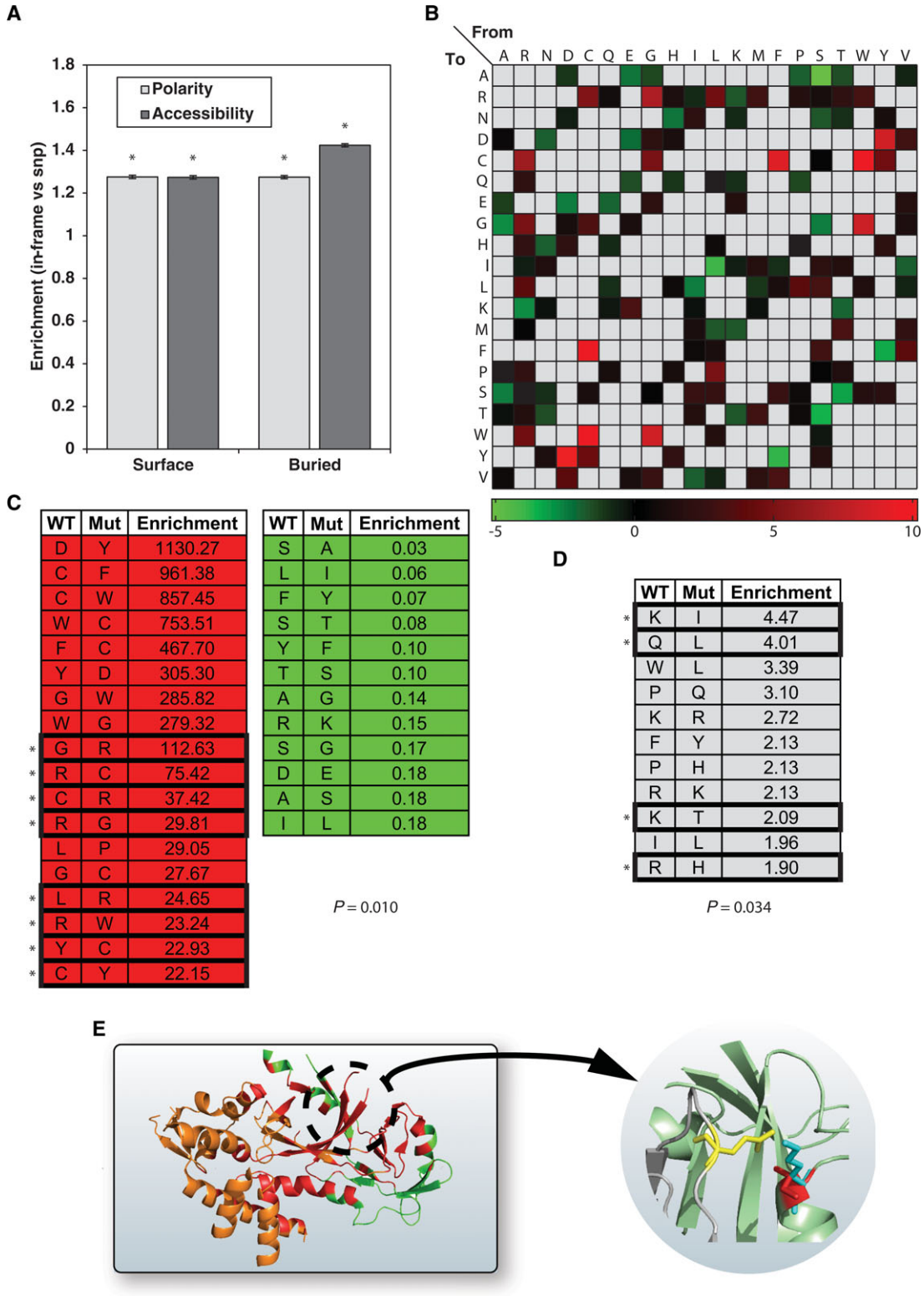


Figure 3. Loci of disease variations associated with interaction strength. **A**: Coexpression profiles for interactions with low, medium, and high buried surface areas. **B**: Enrichment of interactions with low, medium, and high buried surface areas in stable complexes. **C**: Odds ratio for the distribution of in-frame variations at the interface in interactions with low, medium, and high buried surface areas. $^{*}P < 10^{-3}$. **D:** Odds ratio of in-frame variations in the remainder of the interacting domain in interactions with low, medium, and high buried surface areas. (E) Odds ratio of in-frame variations in the rest of the protein in interactions with low, medium, and high buried surface areas. $^{*}P < 10^{-3}$. Error bars indicate ±SE.

($\rho = 0.63$, $P < 10^{-3}$ using a permutation test) between K_a and buried surface area, confirming that the latter is an appropriate surrogate for interaction strength.

Next, we determined the distribution of in-frame variations in different parts of the protein as a function of the strength of the interaction. We found that variations at the interface tend to disrupt strong interactions (odds ratio = 1.10, $P = 0.005$), whereas those in the rest of the protein outside the interacting domains tend to be enriched in weak interactions (odds ratio = 1.24, $P < 10^{-3}$; Figs. 3C–E). As a control, we also computed the distribution of SNPs in different parts of the protein as a function of interaction strength. We found that SNPs at the interface and away from the interface are both randomly distributed with respect to interaction strength (Supp. Fig. S5). Our results therefore suggest that there is a relationship between the location of the disease variation and the biophysical strength of the interactions it disrupts. Because pathogenic variations are enriched at the interaction interface and interface variations selectively affect biophysically strong interactions, we surmise that many strong interactions within stable protein complexes involved in key cellular functions are likely to be preferentially disrupted in human disease. This provides a molecular-level biophysical explanation for the results of previous studies that have suggested that protein complexes are useful predictors for discovering unknown disease genes [Fraser and Plotkin, 2007].



WT	Mut	Enrichment
D	Y	1130.27
C	F	961.38
C	W	857.45
W	C	753.51
F	C	467.70
Y	D	305.30
G	W	285.82
W	G	279.32
* G	R	112.63
* R	C	75.42
* C	R	37.42
* R	G	29.81
L	P	29.05
G	C	27.67
* L	R	24.65
* R	W	23.24
* Y	C	22.93
* C	Y	22.15

WT	Mut	Enrichment
S	A	0.03
L	I	0.06
F	Y	0.07
S	T	0.08
Y	F	0.10
T	S	0.10
A	G	0.14
R	K	0.15
S	G	0.17
D	E	0.18
A	S	0.18
I	L	0.18

WT	Mut	Enrichment
* K	I	4.47
* Q	L	4.01
W	L	3.39
P	Q	3.10
K	R	2.72
F	Y	2.13
P	H	2.13
R	K	2.13
* K	T	2.09
I	L	1.96
* R	H	1.90

Figure 4. Alterations of biochemical properties of individual amino acids in disease. **A:** Enrichment of disease variations that alter the structural (accessibility) and biochemical (polarity) properties of amino acids as compared with SNPs. $^{*}P < 10^{-3}$. Error bars indicate ±SE. **B**: Relative enrichment of all pairs of amino acid changes in human pathological variations as compared with changes that occurred, and that were fixed, during the course of evolution (gray indicates that these pathogenic variations are not observed). **C:** Pairs of amino acid changes enriched in pathogenic missense variations and changes that occurred during evolution (highlighted pairs undergo significant change in biochemical properties). **D**: Pairs of amino acid changes enriched at the atomic-resolution interaction interface (highlighted pairs undergo significant change in biochemical properties). **E**: An example of the alteration of the interaction interface between RNASEH2B and RNASEH2C by a variation (K143I) in RNASEH2C that significantly alters biochemical properties (in the circular panel, the blue residue is the wild-type K and the red residue is the pathogenic variant I).

Significant Alterations in Structural and Biochemical Properties of Amino Acids Involved in Human-Inherited Disease

To systematically explore the structural properties of human pathogenic variations, we analyzed relationships between the properties of these variations and their accessibility in the protein. Amino acids may be classified as either accessible or inaccessible (Supp. Note S4 and Supp. Fig. S6). Using the Janin accessibility scale [Janin, 1979], we then calculated the proportion of accessibility altering in-frame missense disease-associated variations (i.e., point variations that cause an accessible wild-type amino acid to be changed to an inaccessible amino acid or vice versa) in different parts of the protein. These variations are most likely to cause dramatic changes to the configuration of the interface because the local structural configuration is drastically altered. Since disease-associated variations in different parts of the protein may exert their effects via different pathophysiological mechanisms, we normalized our results by calculating the ratio of accessibility altering in-frame variations against a background distribution of putatively neutral SNPs that are characterized by a similar change in their accessibility. Since these SNPs are uniformly distributed throughout the protein (Fig. 1B), this gives us an idea of the relative propensity of disease-associated variations to be significantly accessibility altering. We found that at both surface and buried residues, and indeed in all parts of the protein, accessibility altering variations are significantly more likely to occur in pathogenic variations as opposed to putatively neutral variants in the general population ($P < 10^{-3}$; Fig. 4A).

We also examined amino acid substitutions in terms of their change in polarity (Supp. Note S5 and Supp. Fig. S6). We calculated the proportion of polarity altering in-frame missense disease-associated variations (i.e., those that cause a polar wild-type amino acid to change to a nonpolar amino acid or vice versa). We note that these alterations also follow a similar trend—at both surface and buried residues, and in all regions of the protein, polarity altering variations are significantly more likely to occur in disease as opposed to putatively neutral variants in the population ($P < 10^{-3}$; Fig. 4A). This suggests that disease-associated variations are biochemically more destabilizing to the protein than benign variants in the population.

To further understand how disease-associated variations differ in terms of their biochemical properties from changes that have been fixed over the course of evolutionary time, we calculated the relative enrichment of all possible pairs of amino acid changes for disease-associated in-frame missense variations over those that have occurred during evolution. We obtained the probabilities of amino acid changes occurring during evolution from a recently updated version of the Dayhoff matrices [Kosiol and Goldman, 2005]. We compared these amino acid changes to in-frame disease variations occurring throughput the protein (Fig. 4B). We found that disease-associated variations generally tend to alter accessibility of the wild-type amino acid, whereas evolutionary changes tend to preserve it ($P = 0.010$; Fig. 4C, Supp. Note S6 and Supp. Fig. S7). Our findings contrast with previous reports of significant correlations between amino acid variations in genetic disease and evolution [Wu et al., 2007]. To further understand the specific differences in the distribution of variations in different parts of the protein, we determined which variations were enriched at least twofold at the interaction interface compared with other regions of the protein (Fig. 4D). We found that these interface variations are significantly more likely to change the accessibility of the amino acid involved ($P = 0.034$), with the most dramatic changes occurring with those variations with the highest enrichment (Supp. Note S7 and Supp. Fig. S8).

By way of an example, a K143I variation at the interaction interface of RNASEH2B and RNASEH2C has been shown to be associated with a human autoinflammatory disorder, Aicardi–Goutières syndrome [Reijns et al., 2011]. This variation causes a major change in structural and biochemical properties, leading to a significant structural modification at the interface that specifically alters the wild-type interaction (Fig. 4E). These results further validate our finding that pathogenic variations tend to be more disruptive than random evolutionary changes, with those occurring at the protein interface causing the most drastic changes, enough to perturb even strong interactions.

In this study, we build and use an atomic-resolution human protein interactome network to improve our understanding of the structural principles and molecular mechanisms of pathogenic variations that perturb protein–protein interactions leading to disease. We find that in-frame variations are significantly enriched both at the interaction interface as well as in the remainder of the corresponding interacting domain. Thus, it is not just the residues at the interface that serve as the key mediators of interactions [Jones and Thornton, 1996; Hu et al., 2000], variations outside the interface but within the interacting domain are capable of altering protein–protein interactions. Our findings suggest that it is the alteration of specific interactions by in-frame variations within the entire interacting domain that is a major molecular determinant of human-inherited disease. Moreover, we show that there are important biochemical and biophysical differences between variations at the interface and those located in the remainder of the protein molecule. Specifically, we find that the locations of pathogenic variations are associated with the strength of interactions—those at the interface tend to selectively disrupt stronger interactions. One mechanistic explanation for such a phenomenon is the tendency for variations enriched at the interface (as compared with other parts of the protein) to cause the most dramatic changes in their structural and biochemical properties. Analyses at the level of individual amino acids are only possible with atomic-resolution interactome networks. Our findings suggest that the structurally guided prioritization of pathogenic variations identified in large-scale sequencing studies using an atomic-resolution network might be useful in the context of informing follow-up experiments.

The coverage of the atomic-resolution human protein interactome network is limited by the number of cocrystal structures currently available in PDB. As more cocrystal structures become available [Chandonia and Brenner, 2006], the same principles developed here can be readily applied to reveal additional specific structural mechanisms underlying pathogenic variations. Our work further underscores the importance of the exploration of all possible domain architectures by structural genomics consortia [Editorial, 2007]. Using our methodology on a more complete set of structural folds is likely to generate reliable direct atomic-resolution target sites for structurally aided rational drug design, and has the potential to overcome the difficulties routinely encountered due to the paucity of well-elucidated structural targets [Tanrikulu and Schneider, 2008; Xie and Bourne, 2011].

Acknowledgment

The authors would like to thank Sandipan Chowdhury for insightful discussions regarding the manuscript.

Disclosure statement: The authors declare no conflict of interest.

References

Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. 2008. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36:D419–D425.

Assi SA, Tanaka T, Rabbitts TH, Fernandez-Fuentes N. 2010. PCRPi: presaging critical residues in protein interfaces, a new computational tool to chart hot spots in protein interfaces. Nucleic Acids Res 38:e86.

Bao L, Cui Y. 2005. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics 21:2185–2190.

Barabasi AL, Gulbahce N, Loscalzo J. 2011. Network medicine: a network-based approach to human disease. Nat Rev Genet 12:56–68.

Behrends C, Sowa ME, Gygi SP, Harper JW. 2010. Network organization of the human autophagy system. Nature 466:68–76.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Res 28:235–242.

Celli J, Duijf P, Hamel BC, Bamshad M, Kramer B, Smits AP, Newbury-Ecob R, Hennekam RC, Van Buggenhout G, van Haeringen A, Woods CG, van Essen AJ, et al. 1999. Heterozygous germline mutations in the p53 homolog p63 are the cause of EEC syndrome. Cell 99:143–153.

Chandonia JM, Brenner SE. 2006. The impact of structural genomics: expectations and outcomes. Science 311:347–351.

Chayen NE, Saridakis E. 2008. Protein crystallization: from purified protein to diffraction-quality crystal. Nat Methods 5:147–153.

David A, Razali R, Wass MN, Sternberg MJ. 2012. Protein–protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. Hum Mutat 33:359–363.

de Souza N. 2012. Systems biology: a bird's-eye view of disease. Nat Meth 9:220–221.

Editorial. 2007. Looking ahead with structural genomics. Nat Struct Mol Biol 14:1.

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

Franzosa EA, Xia Y. 2011. Structural principles within the human-virus protein-protein interaction network. Proc Natl Acad Sci USA 108:10538–10543.

Fraser HB, Plotkin JB. 2007. Using protein complexes to predict phenotypic effects of gene mutation. Genome Biol 8:R252.

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, Nickerson DA, Bamshad MJ, et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493:216–220.

Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A. 1999. Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations. Protein Eng 12:549–555.

Haberle J, Shchelochkov OA, Wang J, Katsonis P, Hall L, Reiss S, Eeds A, Willis A, Yadav M, Summar S, Lichtarge O, Rubio V, et al. 2011. Molecular defects in human carbamoy phosphate synthetase I: mutational spectrum, diagnostic and protein structure considerations. Hum Mutat 32:579–589.

Hu Z, Ma B, Wolfson H, Nussinov R. 2000. Conservation of polar residues as hot spots at protein interfaces. Proteins 39:331–342.

Hubbard SJ, Thornton JM. 1993. 'NACCESS', computer program.

Janin J. 1979. Surface and inside volumes in globular proteins. Nature 277:491–492.

Jones S, Thornton JM. 1996. Principles of protein–protein interactions. Proc Natl Acad Sci USA 93:13–20.

Khurana E, Fu Y, Chen J, Gerstein M. 2013. Interpretation of genomic variants using a unified biological network approach. PLoS Comput Biol 9:e1002886.

Kosiol C, Goldman N. 2005. Different versions of the Dayhoff rate matrix. Mol Biol Evol 22:193–199.

Lahiry P, Torkamani A, Schork NJ, Hegele RA. 2010. Kinase mutations in human disease: interpreting genotype–phenotype relationships. Nat Rev Genet 11:60–74.

Marks DS, Hopf TA, Sander C. 2012. Protein structure prediction from sequence variation. Nat Biotechnol 30:1072–1080.

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

Moal IH, Fernandez-Recio J. 2012. SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. Bioinformatics 28:2600–2607.

Noskov SY, Lim C. 2001. Free energy decomposition of protein–protein interactions. Biophys J 81:737–750.

Oppliger T, Thony B, Nar H, Burgisser D, Huber R, Heizmann CW, Blau N. 1995. Structural and functional consequences of mutations in 6-pyruvoyltetrahydropterin synthase causing hyperphenylalaninemia in humans. Phosphorylation is a requirement for in vivo activity. J Biol Chem 270:29498–29506.

Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D, Akpor A, Maibaum M, et al. 2005. The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res 33:D247–D251.

Reijns MA, Bubeck D, Gibson LC, Graham SC, Baillie GS, Jones EY, Jackson AP. 2011. The structure of the human RNase H2 complex defines key interaction interfaces relevant to enzyme function and human disease. J Biol Chem 286:10530–10539.

Shi YY, Miller GA, Qian H, Bomsztyk K. 2006. Free-energy distribution of binary protein–protein binding suggests cross-species interactome differences. Proc Natl Acad Sci USA 103:11527–11532.

Stein A, Ceol A, Aloy P. 2011. 3did: identification and classification of domain-based interactions of known three-dimensional structure. Nucleic Acids Res 39:D718–D723.

Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, Cooper DN. 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133:1–9.

Studer RA, Dessailly BH, Orengo CA. 2013. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. Biochem J 449:581–594.

Sunyaev S, Ramensky V, Koch I, Lathe W 3rd, Kondrashov AS, Bork P. 2001. Prediction of deleterious human alleles. Hum Mol Genet 10:591–597.

Tanrikulu Y, Schneider G. 2008. Pseudoreceptor models in drug design: bridging ligand- and receptor-based virtual screening. Nat Rev Drug Discov 7:667–677.

Thony B, Leimbacher W, Blau N, Harvie A, Heizmann CW. 1994. Hyperphenylalaninemia due to defects in tetrahydrobiopterin metabolism: molecular characterization of mutations in 6-pyruvoyl-tetrahydropterin synthase. Am J Hum Genet 54:782–792.

Tuncbag N, Gursoy A, Nussinov R, Keskin O. 2011. Predicting protein–protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. Nat Protoc 6:1341–1354.

Vidal M. 2005. Interactome modeling. FEBS Lett 579:1834–1838.

Vidal M, Cusick ME, Barabasi AL. 2011. Interactome networks and human disease. Cell 144:986–998.

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. 2002. Comparative assessment of large-scale data sets of protein–protein interactions. Nature 417:399–403.

Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H. 2012. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol 30:159–164.

Wells JA, McClendon CL. 2007. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. Nature 450:1001–1009.

Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO, Botstein D. 2002. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell 13:1977–2000.

Wu H, Ma BG, Zhao JT, Zhang HY. 2007. How similar are amino acid mutations in human genetic diseases and evolution. Biochem Biophys Res Commun 362:233–237.

Xie L, Bourne PE. 2011. Structure-based systems biology for analyzing off-target binding. Curr Opin Struct Biol 21:189–199.

Yang X, Boehm JS, Salehi-Ashtiani K, Hao T, Shen Y, Lubonja R, Thomas SR, Alkan O, Bhimdi T, Green TM, Johannessen CM, Silver SJ, et al. 2011. A public genome-scale lentiviral expression library of human ORFs. Nat Methods 8:659–661.

Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, Hao T, Rual JF, et al. 2008. High-quality binary protein interaction map of the yeast interactome network. Science 322:104–110.

Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Maniatis T, Califano A, et al. 2012. Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature 490:556–560.

Zhong Q, Simonis N, Li QR, Charloteaux B, Heuze F, Klitgord N, Tam S, Yu H, Venkatesan K, Mou D, Swearingen V, Yildirim MA, et al. 2009. Edgetic perturbation models of human inherited disorders. Mol Syst Biol 5:321.