Molecular Systems Biology 5; Article number 321; doi:10.1038/msb.2009.80
Citation: ***Molecular Systems Biology*** **5:321**
© 2009 EMBO and Macmillan Publishers Limited All rights reserved 1744-4292/09
www.molecularsystemsbiology.com

Edgetic perturbation models of human inherited disorders

Quan Zhong[1,2,4], Nicolas Simonis[1,2,4], Qian-Ru Li[1,2,4], Benoit Charloteaux[1,2,3,4], Fabien Heuze[1,2,3,4], Niels Klitgord[1,2,4], Stanley Tam[1,2], Haiyuan Yu[1,2], Kavitha Venkatesan[1,2], Danny Mou[1,2], Venus Swearingen[1,2], Muhammed A Yildirim[1,2], Han Yan[1,2], Amélie Dricot[1,2], David Szeto[1,2], Chenwei Lin[1,2], Tong Hao[1,2], Changyu Fan[1,2], Stuart Milstein[1,2], Denis Dupuy[1,2], Robert Brasseur[3], David E Hill[1,2], Michael E Cusick[1,2] and Marc Vidal[1,2,]*

[1] Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA, [2] Department of Genetics, Harvard Medical School, Boston, MA, USA and [3] Centre de Biophysique Moléculaire Numérique, Faculté Universitaire des Sciences Agronomiques de Gembloux, Gembloux, Wallonia, Belgium

[4] These authors contributed equally to this work

* Corresponding author. Center for Cancer Systems Biology, Department of Cancer Biology, 44 Binney Street, Smith 858, Boston, MA 02115, USA. Tel.: +1 617 632 5180; Fax: +1 617 632 5739; E-mail: marc_vidal@dfci.harvard.edu

Received 5.8.09; accepted 2.10.09

Cellular functions are mediated through complex systems of macromolecules and metabolites linked through biochemical and physical interactions, represented in interactome models as 'nodes' and 'edges', respectively. Better understanding of genotype-to-phenotype relationships in human disease will require modeling of how disease-causing mutations affect systems or interactome properties. Here we investigate how perturbations of interactome networks may differ between complete loss of gene products ('node removal') and interaction-specific or edge-specific ('edgetic') alterations. Global computational analyses of ∼50 000 known causative mutations in human Mendelian disorders revealed clear separations of mutations probably corresponding to those of node removal versus edgetic perturbations. Experimental characterization of mutant alleles in various disorders identified diverse edgetic interaction profiles of mutant proteins, which correlated with distinct structural properties of disease proteins and disease mechanisms. Edgetic perturbations seem to confer distinct functional consequences from node removal because a large fraction of cases in which a single gene is linked to multiple disorders can be modeled by distinguishing edgetic network perturbations. Edgetic network perturbation models might improve both the understanding of dissemination of disease alleles in human populations and the development of molecular therapeutic strategies.

Molecular Systems Biology **5**: 321; published online 3 November 2009; doi:10.1038/msb.2009.80
Subject Categories: molecular biology of disease
Keywords: binary protein interaction; genotype-to-phenotype relationships; human Mendelian disorders; network perturbation

Introduction

Decades of research into human Mendelian disorders has led to the discovery of a massive amount of disease-associated allelic variations. Most disease-causing mutations are thought to confer radical changes to proteins (Wang and Moult, 2001; Botstein and Risch, 2003; Yue *et al*, 2005; Subramanian and Kumar, 2006). Consequently, genotype-to-phenotype relationships in human genetic disorders are often modeled as: 'mutation in gene *X* leads to loss of gene product X, which leads to disease A'. A single 'gene-loss' model seems pertinent for many diseases (Botstein and Risch, 2003). However, this model cannot fully reconcile with the increasingly appreciated prevalence of complex genotype-to-phenotype associations for even 'simple' Mendelian disorders (Goh *et al*, 2007), particularly in which: (i) a single gene can be associated with multiple disorders (allelic heterogeneity), (ii) a single disorder can be caused by mutations in any one of several genes (locus heterogeneity), (iii) only a subset of individuals carrying a mutation are affected by the disease (incomplete penetrance), or (iv) not all individuals with a given mutation are affected equally (variable expressivity). More complex models to interpret genotype-to-phenotype relationships would probably improve the understanding of human disease.

Genes and gene products function not in isolation but as components of complex networks of macromolecules (DNA, RNA, or proteins) and metabolites linked through biochemical or physical interactions, often represented in 'interactome' network models as 'nodes' and 'edges', respectively. Cellular networks seem to exhibit systems properties underlying phenotypic variations (Goh *et al*, 2007). Here we propose network-perturbation models to explain molecular dysfunctions underlying human disease.

We hypothesize that distinct mutations causing distinct molecular defects to proteins may lead to distinct perturbations of cellular networks, giving rise to distinct phenotypic outcomes (Figure 1A). Truncations close to the start of an open-reading frame, or mutations that grossly destabilize a protein structure, can be modeled as removing a protein node from the network ('node removal'). Alternatively, single amino-acid substitutions that affect specific binding sites, or truncations that preserve certain domains of a protein, may give rise to partially functional gene products with specific changes in distinct biophysical or biochemical interaction(s) (edge-specific genetic perturbation or 'edgetic' perturbations; Figure 1B).

Edgetic network perturbations provide alternative molecular explanations for protein dysfunction in addition to gene loss. Taking advantage of the large number of known disease-causing allelic variations in human Mendelian disorders, we investigated how such mutations may cause complete loss of gene products or, alternatively, cause specific loss or gain of distinct molecular interaction(s). We further tested edgetic perturbation models in cases in which a single gene is associated with multiple disorders. Together, both experimental and computational evidence support edgetic perturbation models in human inherited disorders. Edgetic perturbations probably underlie many complex genotype-to-phenotype relationships.

Results

Global distribution of disease-causing mutations

To investigate possibly differing network perturbations in human inherited disorders, we examined ∼50 000 Mendelian disease-causing alleles, affecting over 1900 protein-coding genes, altogether associated with more than 2000 human disorders available in the Human Gene Mutation Database (HGMD) (Stenson *et al*, 2003). We differentiated all disease alleles into two subsets probably causing different molecular defects to proteins. The first subset ('truncating' alleles) comprises all mutations that lead to the synthesis of truncated gene products, including nonsense mutations, out-of-frame insertions or deletions, or defective splicing. The second subset ('in-frame' alleles) comprises mutations that probably give rise to nearly full-length gene products, including missense mutations and in-frame insertions or deletions. Over 50% (27 919/52 491) of Mendelian alleles in HGMD correspond to 'in-frame' alleles (Figure 2A). Our hypothesis is that 'truncating' and 'in-frame' alleles probably cause distinct

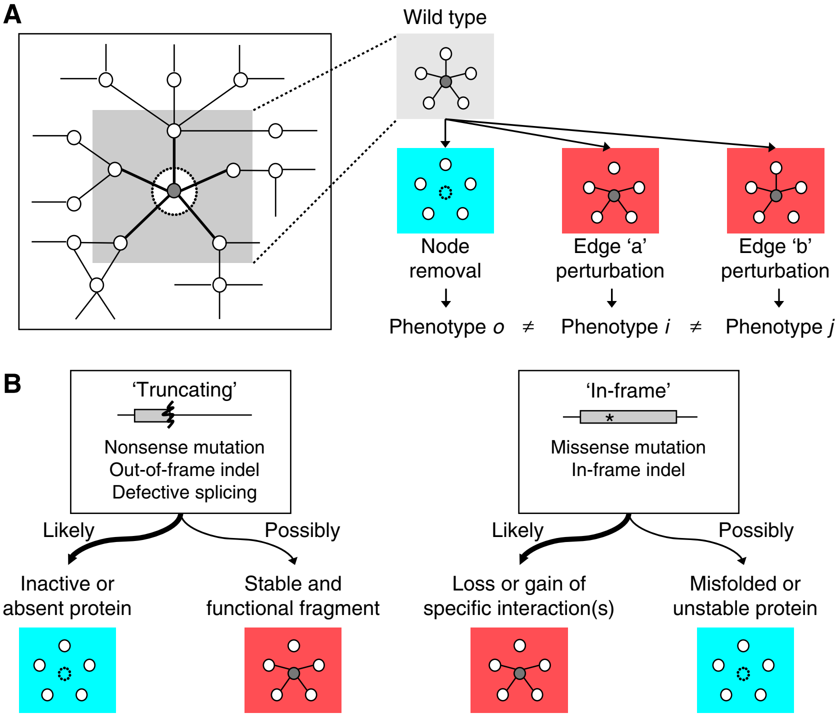

Figure 1 Node removal versus edgetic perturbation models of network changes underlying phenotypic alterations. (**A**) Schematic illustration of pleiotropic phenotypic outcomes resulting from distinct network perturbations upon complete loss of gene product (node removal, blue box) versus perturbation of specific molecular interactions (edgetic perturbation, red box). Solid lines between two nodes represent preserved interactions and dashed lines represent perturbed interactions. Edges are generally biophysical interactions, but could also be biochemical interactions. (**B**) Schematic illustration of distinct 'truncating' versus 'in-frame' mutations causing distinct molecular defects in proteins leading to distinct node removal versus edgetic perturbation.

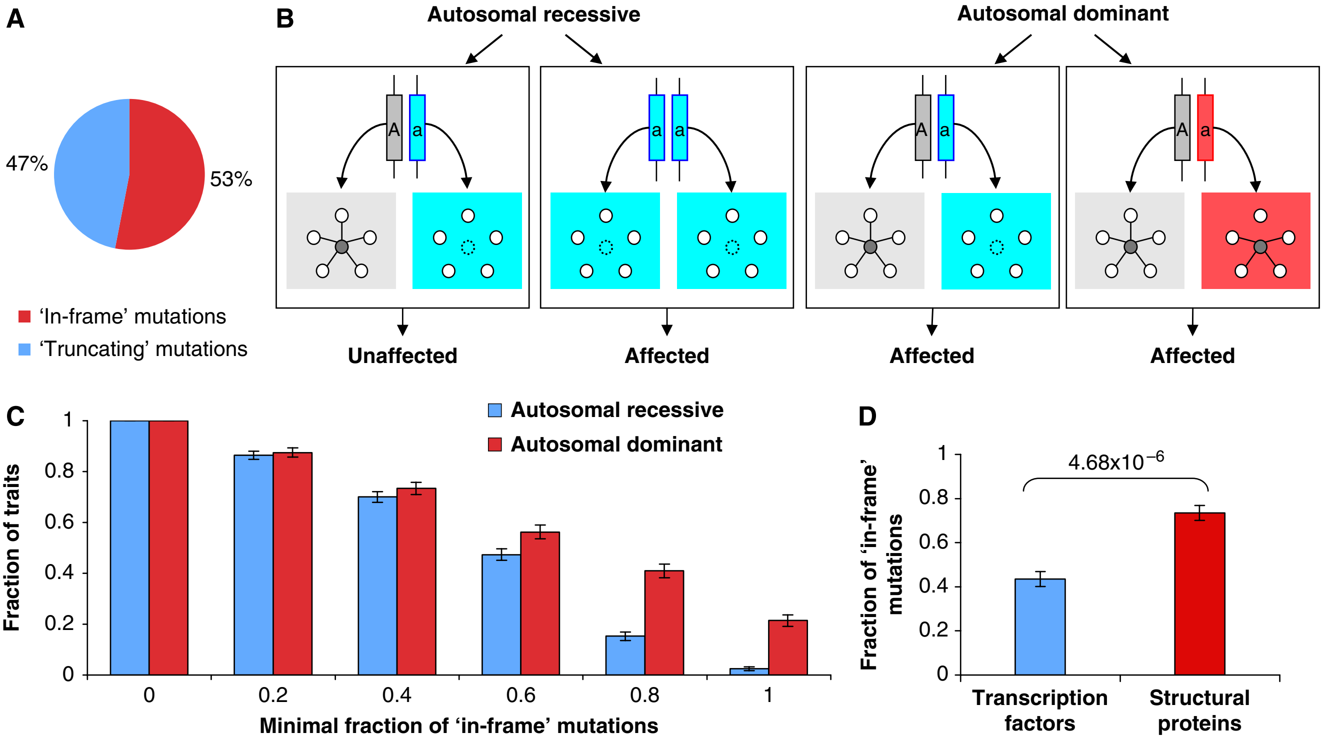

Figure 2 Global patterns of disease mutations in human genetic disorders. (**A**) Subdivision of 'truncating' versus 'in-frame' mutations in Human Gene Mutation Database (HGMD) (Stenson *et al*, 2003). (**B**) Schematic illustration of distinct node removal versus edgetic perturbation models in disease with autosomal recessive versus autosomal dominant inheritance. (**C**) Distribution of autosomal recessive and dominant disease with respect to the associated 'in-frame' mutations. Mutations in each gene associated with each mode of inheritance are grouped as one trait. Each data point represents the fraction of autosomal recessive (blue bar) or autosomal dominant (red bar) traits that have a fraction of 'in-frame' mutations no less than the value on the *x*-axis. Statistical significance of the observed difference between distributions is assessed by Mann–Whitney *U* test ($P<9.2\times10^{-12}$). The number of traits, genes, diseases and total mutations in each bin are provided in Supplementary Table 1. (**D**) Average fraction of 'in-frame' mutations associated with autosomal dominant disease in transcription factors and structural proteins. *P*-value assessed by Mann–Whitney *U* test of the observed difference is shown.

molecular defects in proteins, and are thus enriched in distinct node removal or edgetic perturbations, respectively. This hypothesis is based on the assumption that 'truncating' alleles are less prone to produce stably folded proteins than 'in-frame' alleles. Although exceptions may apply, our hypothesis predicts that 'truncating' versus 'in-frame' alleles may distribute differently among diseases involving distinct node removal versus edgetic perturbations.

Given that, with the exception of haploinsufficiency, many established molecular explanations for dominance entail production of a mutated protein that interferes in some way with the function of the product of the normal allele, autosomal dominant disease should be more frequently associated with edgetic perturbation than node removal (Figure 2B). To test the hypothesis that 'truncating' versus 'in-frame' alleles are enriched in distinct node removal versus edgetic perturbations, respectively, we retrieved the inheritance information, by manual curation, for each HGMD-annotated phenotype from the Online Mendelian Inheritance in Man (OMIM) database (Hamosh *et al*, 2005). 'Truncating' versus 'in-frame' alleles distribute differently among autosomal dominant and autosomal recessive traits. Among genes affected solely by 'in-frame' mutations, the proportion of autosomal dominant diseases is ~10-fold higher than that of autosomal recessive diseases (Figure 2C). This trend holds even after removing all human predicted orthologs of essential genes from the analysis (Supplementary Figure S1).

We next examined whether distinct distribution of 'truncating' versus 'in-frame' alleles can also be found among autosomal dominant traits that are probably caused by different molecular mechanisms. Mutations in cytoskeleton proteins frequently cause dominant-negative effects, in which incorporation of expressed abnormal molecules into multimeric assemblies of structural proteins disrupts the integrity and function of the complex (Wilkie, 1994). In contrast, germline mutations in transcription factors are more frequently associated with haploinsufficiency (Wilkie, 1994; Seidman and Seidman, 2002) probably because of insufficient activity or production of the remaining wild-type allele in heterozygotes. Consistent with this distinction, a significantly higher fraction of 'in-frame' mutations was found for autosomal dominant Mendelian disorders associated with structural proteins than with transcription factors (Figure 2D).

Distinct global distributions of 'truncating' versus 'in-frame' mutations among diseases with distinct modes of inheritance, and in proteins probably associated with distinct molecular mechanisms of dominance, support our hypothesis that 'truncating' versus 'in-frame' alleles are probably enriched in distinct node removal versus edgetic perturbations, respectively. The distinctions observed between autosomal dominant and autosomal recessive mutations may be more pronounced if haploinsufficiency could be separated overall from dominant-negative and other molecular mechanisms of dominance, but such information is currently unavailable at the global level.

Distinguishing edgetic perturbation from node removal

For a proof-of-principle analysis of allele-specific network perturbations by disease proteins, we used an integrated experimental approach to characterize binary protein interaction defects of disease-causing mutant alleles. Our approach includes (i) Gateway recombinational cloning of mutations by PCR-based site-directed mutagenesis (Suzuki *et al*, 2005), (ii) high-throughput mapping of binary protein–protein interactions (Rual *et al*, 2005), (iii) high-throughput characterization of protein–protein interaction defects of all cloned disease-causing mutant proteins, and (iv) integration of network perturbations by disease-causing mutations with structural or functional information of disease proteins.

We selected disease proteins that have: (i) multiple mutations annotated in HGMD (Stenson *et al*, 2003), (ii) wild-type clones available in our human ORFeome collection, hORFeome 3.1 (Lamesch *et al*, 2007), (iii) structural information available in Protein Data Bank (PDB, http://www.rcsb.org/pdb), and (iv) two or more interactions reported in our previous binary human interactome map (Rual *et al*, 2005). We also requested that at least one of the observed interactions by yeast two-hybrid (Y2H) analysis be supported by functional characterization in the literature. Given these criteria, we could apply our allele-profiling platform to one autosomal recessive disease protein (CBS), and to three autosomal dominant disease proteins with likely dominant-negative (ACTG1), abnormal activation (CDK4), or haploinsufficiency (PRKAR1A) molecular defects (Figure 3A). We included one additional autosomal recessive disease protein (HGD) that meets all criteria except that no protein–protein interaction data were available (Figure 3A). We carried out a genome-wide Y2H screen against a set of ~8100 human open-reading frames (Rual *et al*, 2005), and identified three interactions for wild-type HGD. We cloned disease-causing mutants annotated in HGMD for these five proteins and profiled each mutant against the corresponding wild-type interactors.

Profiling interaction defects of 29 alleles associated with five distinct genetic disorders revealed three classes of interaction-defective alleles (Supplementary information and Figure 3B): (i) five alleles that behaved as null, eliminating all interactions, (ii) 16 edgetic alleles that lost specific interaction(s) while retaining other interactions, and (iii) eight alleles that behaved as 'pseudo-wild-type', retaining all currently available protein–protein interactions tested here. Null-like alleles were observed only for two autosomal recessive disease proteins (CBS and HGD) and in a supposed case of dominant haploinsufficiency (PRKAR1A), consistent with differing network perturbations in diseases associated with distinct modes of inheritance (Figure 2B). We propose that many disease-causing alleles scoring as pseudo-wild-type in the assay described here might still be true edgetic alleles. Further analysis with additional physical and biochemical interactors using additional assays should eventually settle that question.

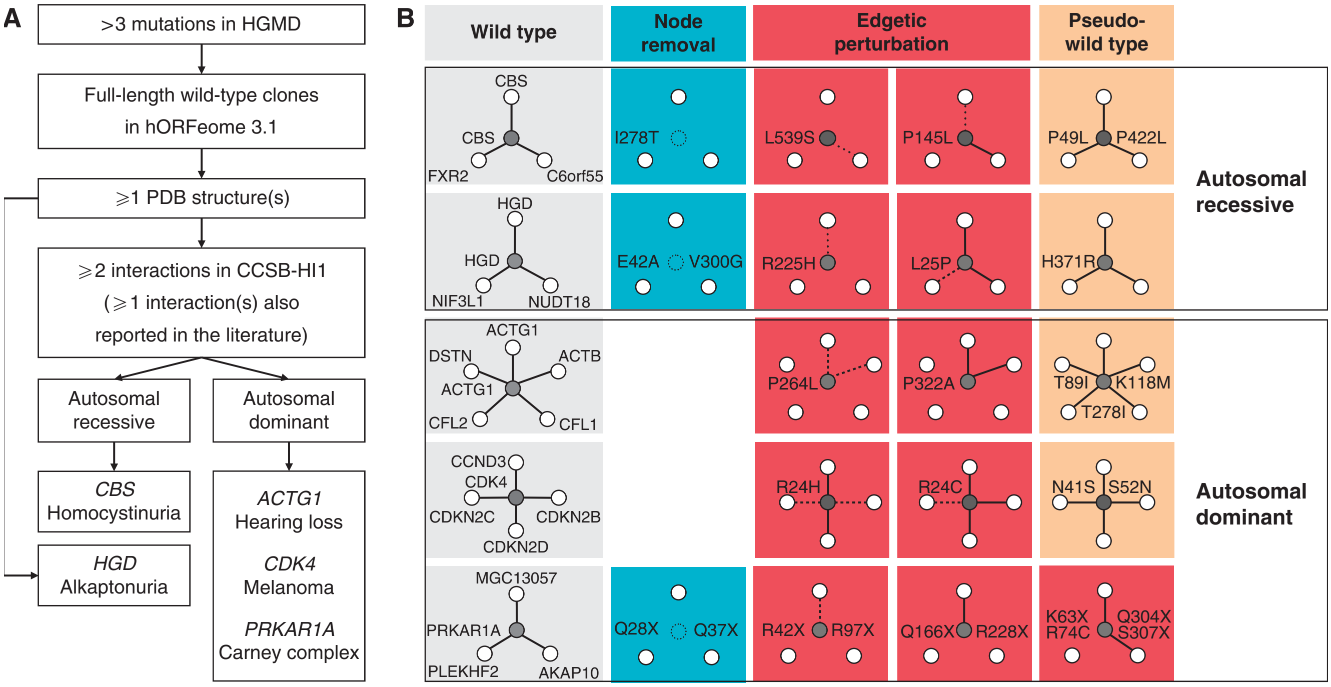

Figure 3 Profiling allele-specific interaction defects of disease-causing mutant proteins. (**A**) Schematic illustration of selection of disease proteins for proof-of-principle analysis of binary protein interaction defects of disease-causing mutant proteins. (**B**) Interpreted network perturbations for each allele comparing to corresponding wild-type proteins. Missing lines represent lost protein interactions. Dashed lines represent reduced protein interactions. Color codes for distinct network perturbations are indicated at the top panel.

We related Y2H interaction profiles of each mutant to structural properties of disease proteins (Supplementary information and Supplementary Figure S2–6). Grossly disruptive mutations tend to affect buried residues of the protein, whereas mutations leading to loss or gain of specific interaction(s) tend to lie on the surface. Edgetic perturbation of some disease alleles revealed diverse molecular mechanisms of protein dysfunction (Supplementary information). Complex allele-specific perturbations were also found to be associated with phenotypic variability among patients, such as their response to specific treatments (Supplementary information for CBS).

Structural analyses of disease-causing mutations

To further investigate the extent to which mutations found in human genetic disorders may grossly disrupt proteins or cause alterations in specific biochemical or biophysical interaction(s), we examined available three-dimensional structures of all disease proteins. As grossly disruptive mutations versus mutations leading to loss or gain of specific interaction(s) probably distribute differently on protein structures (Figure 4A), we divided missense disease-causing mutations into three non-redundant categories: buried residues (<5% of surface accessible to water), exposed residues (⩾30% of surface accessible to water), and residues with intermediate exposure (5–30% of surface accessible to water). Among all 3664 affected residues in 236 proteins for which three-dimensional X-ray structures are available, about one-third of the mutated residues are buried, whereas another one-third are exposed, probably representing complete loss of gene products versus loss or gain of specific molecular interaction(s), respectively (Supplementary Figure S7). Consistent with differing network perturbations in disease with distinct modes of inheritance (Figure 2B), autosomal dominant versus autosomal recessive disease mutations exhibit significant separation with respect to their solvent-accessible surface areas ($P<3\times10^{-10}$; Figure 4B). About 40% of mutated residues in autosomal dominant disease are exposed (with relative solvent-accessible surface areas ⩾30%), whereas only 27% of mutated residues in autosomal recessive disease fall in the same category (Figure 4B).

Allele-specific perturbations observed in PRKAR1A (Supplementary Figure S6) indicate that interaction-specific perturbation by truncations is also possible. As ‘truncating’ alleles outside of protein domains may preserve function of certain domains, giving rise to interaction-specific perturbations (Figure 4C), we determined the distribution of ‘truncating’ mutations in Pfam domains (Finn *et al*, 2006). Although disease-causing ‘truncating’ mutations seem to exhibit a random distribution with respect to Pfam domains (enrichment: 1.0, P=0.2), ‘truncating’ mutations in autosomal dominant disease are slightly depleted in Pfam domains, whereas ‘truncating’ mutations in autosomal recessive disease are slightly enriched in Pfam domains (Figure 4D). This finding is consistent with the hypothesis that different ‘truncating’ mutations may cause distinct node removal versus edgetic perturbations giving rise to disease with distinct modes of inheritance. In agreement with distinct molecular

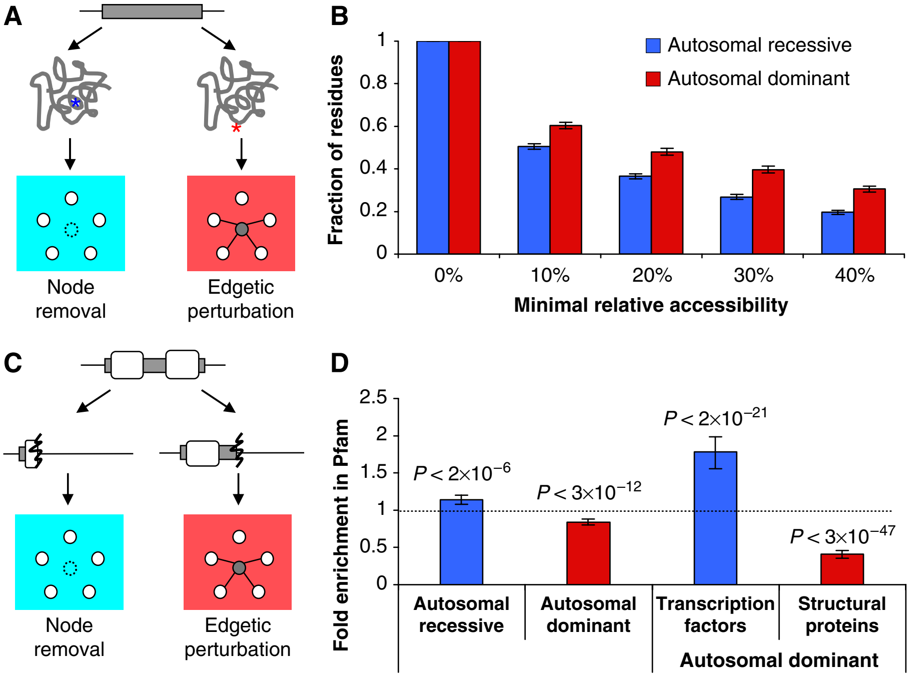

Figure 4 Structural analyses of disease-causing mutations in HGMD. (**A**) Schematic illustration of distinct positions of missense mutations in the three-dimensional structure of a given protein probably causing node removal versus edgetic perturbation. (**B**) Distribution of accessible residues among mutations associated with autosomal recessive diseases (blue bar) and with autosomal dominant ones (red bar). (**C**) Schematic illustration of distinct positions of ‘truncating’ mutations with respect to protein domains probably causing node removal versus edgetic perturbation. (**D**) Distribution of ‘truncating’ mutations in Pfam domains. Fold enrichment higher than one means that Pfam domains contain more mutations than expected at random, whereas enrichment between zero and one means that Pfam domains are depleted in mutations. *P*-values assess the significance of the observed fold enrichment.

mechanisms of dominance (Figure 2B), we found a depletion of autosomal dominant 'truncating' mutations in Pfam domains for structural proteins against an enrichment for transcription factors (Figure 4D), probably associated with dominant-negative effects versus haploinsufficiency, respectively.

Node removal versus edgetic perturbation in complex gene-disease associations

The complex patterns of disease mutations noted so far indicate that a substantial fraction of causative alleles in human genetic disorders may cause edgetic perturbations rather than node removal. Distinct network perturbation models, leading to distinct phenotypic outcomes (Figure 1), predict that 'truncating' versus 'in-frame' alleles for a given gene product might cause different diseases (Figure 5A). We therefore examined 142 genes associated with two or more diseases for which at least five distinct alleles have been reported for each disease. Among 278 disease pairs, each associated with a single one of these 142 genes, we found 88 pairs (~30%) for which the proportion of 'in-frame' versus 'truncating' mutations is significantly different between the two diseases ($P<0.05$; Figure 5B and Supplementary Table 2). A noteworthy example involves the four types (I, II, III, and IV) of osteogenesis imperfecta (OI) with *COL1A1* 'in-frame' mutations causing strikingly more severe phenotypes (in type II, III, or IV) than 'truncating' mutations involved in type I (Hamosh *et al*, 2005; Figure 5B).

Among 34 genes that are linked to both autosomal dominant and autosomal recessive disorders, the fraction of 'in-frame' versus 'truncating' mutations per gene is significantly higher for autosomal dominant mutations than for autosomal recessive ones (Supplementary Figure S8). This finding further supports our hypothesis that distinct 'in-frame' versus 'truncating' mutations probably cause distinct network perturbations giving rise to disease with distinct modes of inheritance (Figure 2).

Edgetic interaction profiles of CBS and PRKAR1A mutant proteins (Figure 3) revealed possible connections between allele-specific interaction defects and differential treatment responses or phenotypic severity among patients (Supplementary information). In addition to clinical variability, edgetic perturbation models also predict that distinct edgetic perturbations for a given gene product might cause phenotypically distinguishable disorders (Figure 6A). We used predicted Pfam domains as surrogates for functional protein domains (Sammut *et al*, 2008), assuming that 'in-frame' mutations located in different Pfam domains probably alter protein functions differently. Among 169 genes associated with two or more diseases and encoding proteins containing at least two Pfam domains, 77 had significant enrichment of 'in-frame' mutations in Pfam domains ($P<0.05$). There were nine proteins with at least two Pfam domains significantly enriched with 'in-frame' mutations ($P<0.05$). For each of the nine proteins, we found a striking pattern of near mutual exclusivity, whereby different Pfam domains seem to be specifically affected in distinct disorders (Figure 6B and Supplementary Table 3). A compelling example is *TP63* (van Bokhoven and Brunner, 2002) in which two clinically distinct developmental disorders, ectrodactyly ectodermal dysplasia (EEC) and ankyloblepharon ectodermal dysplasia (AEC), are caused by mutations in two separate domains, one predicted to bind DNA and the other to mediate protein–protein interaction(s) (Figure 6B). Current information on protein functional domains is incomplete, thus limiting the resolution for distinguishing phenotypes and genotypes. With more detailed structural and biochemical information available, more such allele-specific edgetic phenotype-to-genotype correlations should be uncovered.

Discussion

There are commonalities behind disease mutations that have been discerned, such as disease mutations tend to present at highly conserved regions and to confer radical changes to

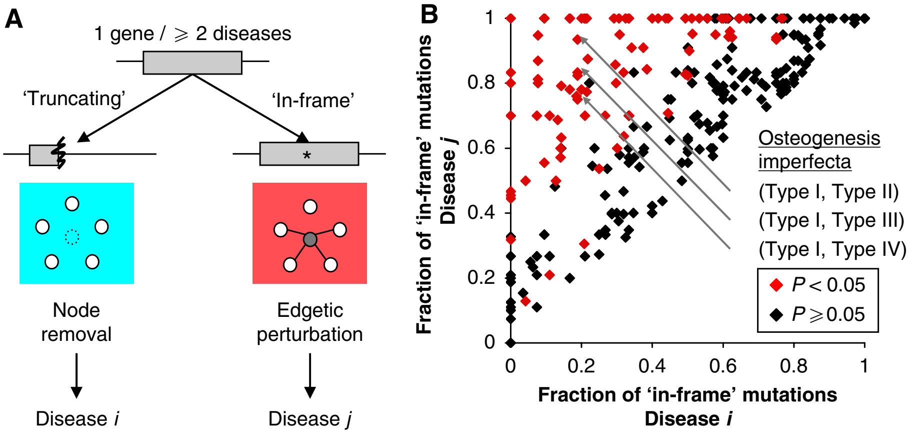

Figure 5 Distinct node removal versus edgetic perturbation underlying pleiotropy. (**A**) Schematic illustration of distinct 'truncating' versus 'in-frame' mutations in a single gene product causing distinct network perturbations giving rise to distinct disorders. (**B**) Analysis of 'in-frame' mutations found in genes associated with multiple diseases. Each dot represents the fraction of 'in-frame' mutations of a pair of distinct diseases associated with a common gene. *x*-axis represents the smaller fraction of 'in-frame' mutation in each pair and *y*-axis represents the larger fraction. Significantly different fractions of 'in-frame' mutation between each pair of diseases are represented by red dots ($P<0.05$). Statistically indistinguishable pairs are represented in black. Three gray arrows pointing to three disease pairs corresponding to Type I and Type II, III or IV Osteogenesis Imperfecta, with 'in-frame' mutation fraction of 0.19 and 0.93, 0.83, 0.75 respectively.

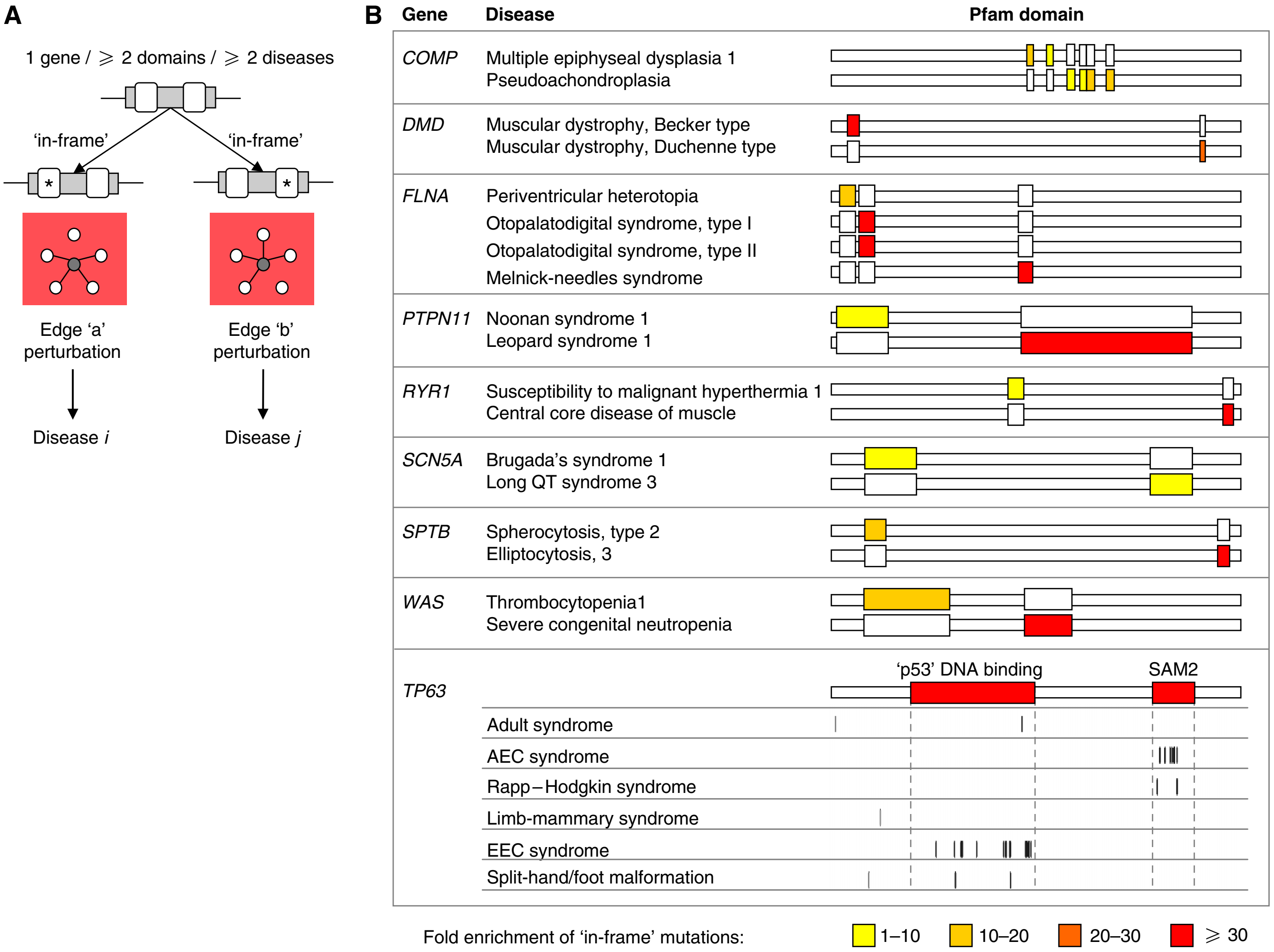

Figure 6 Distinct edgetic perturbations underlying pleiotropy. (**A**) Schematic illustration of distinct 'in-frame' alleles in a single gene product causing distinct network perturbations giving rise to distinct disorders. (**B**) Enrichment of 'in-frame' mutations causing different disorders in different Pfam domains. Color intensity of Pfam domains represents fold enrichment of each disease associated 'in-frame' mutations ($P < 0.05$). Vertical lines below corresponding Pfam domains mark disease-causing 'in-frame' mutations in *TP63*.

proteins (Wang and Moult, 2001; Botstein and Risch, 2003; Yue *et al*, 2005; Subramanian and Kumar, 2006), but there are more complexities to disease mutations and these should not be overlooked. Here we uncovered both experimental and computational evidences that strongly support distinct network perturbations in human Mendelian disorders resulting from complete loss of gene products (node removal) or specific alterations in distinct molecular interaction(s) (edgetic perturbation), respectively (Figures 2–4). Distinct edgetic network perturbations probably underlie many complex genotype-to-phenotype relationships in human genetic disorders (Figures 5 and 6) supporting the idea that edgetic perturbation versus node removal may confer fundamentally different functional consequences.

Edgetic network perturbation models focus on specific alterations in distinct molecular interactions. Although the 'node-centered' gene knockout or knockdown approaches are convenient and useful in determining effects of gross disruption of proteins in model organisms, an 'edge-centered' allele-profiling approach, as carried out here and elsewhere (Dreze *et al*, in press), dissects the dynamics and complexities of biological systems, in which different interactions may occur independently, and in which a single protein may carry out different functions with different partners or in different biological contexts. Edgetic alleles with suboptimal but largely preserved molecular interactions may become insufficient when expressed at reduced levels or may become less stable. Such properties of edgetic alleles may be regulated by other genetic or environmental factors. In this regard, functional characterization of edgetic alleles may help explain phenotypic variations among patients, such as incomplete penetrance or variable expressivity, as well as differential clinical treatment responses (e.g. CBS alleles, Supplementary information). In addition, edgetic network perturbation models might improve our understanding of why and how disease alleles have disseminated in human populations.

Just as high-throughput sequencing technologies are revolutionizing genotyping platforms, and as functional genomics and proteomics are becoming increasingly able to characterize gene products resulting from whole genome sequencing and gene prediction, functional characterizations of genetic variations may be applied at large-scale to characterize mutations with uncertain pathological consequences.

We considered the effects of disease-causing mutations on physical protein–protein interactions, perturbation of which has emerged as a characteristic shared by many disease mutations (Ye *et al*, 2006; Hsu *et al*, 2007; Schuster-Böckler and Bateman, 2008). Complete understanding of network perturbations in disease would require comprehensive analysis of disease mutant proteins by integration of data available from multiple functional assays. First, the current interactome network derived from Y2H analysis is probably incomplete. Many biologically relevant interactors remain to be tested and many may not be recovered by Y2H alone or by any other single protein interaction assay (Braun *et al*, 2009; Venkatesan *et al*, 2009). Second, Y2H detects binary protein interactions. A positive Y2H readout does not necessarily warrant proper protein complex assembly *in vivo*. In oligomer assembly, multiple interaction surfaces of the monomer may be utilized. Mutant alleles that disrupt one but not all interaction surfaces may show positive interaction in the Y2H analysis, but may still affect proper oligomerization. Third, Y2H is not quantitative. Subtle alterations in the affinity of protein–protein interactions, which are undetectable by Y2H, may confer phenotypic changes. Finally, disease mutations may affect protein functions by altering biochemical activities or protein–DNA or protein–RNA interactions.

Disease-associated alleles may also gain new interactions, which is another important potential mechanism for pathogenicity. Gain-of-interaction alleles may be discovered by screening for new interactions specific for an individual mutant. Although we can assay only known edges at any given moment, as more physical and biochemical interactions become identified with time, deeper edgetic profiling will become possible. The pilot step taken here will reach its full potential when applied at genome or proteome scale, with the results integrated into extensive molecular networks.

Materials and methods

Database annotation

The lists of genes and associated phenotypes were downloaded from HGMD website (Stenson *et al*, 2003) (June 2006). The corresponding gene IDs were retrieved from Entrez Gene (Maglott *et al*, 2005) (June 2006). By manual annotation we linked phenotypes associated with each mutation, as annotated in HGMD, to the corresponding disease in the OMIM database (Hamosh *et al*, 2005). The resulting list contains 2269 gene-to-OMIM disease ID entries associated with 48 774 distinct mutations. We carried out all analyses on the resulting gene–OMIM disease associations. We obtained the inheritance information for the corresponding disease available in OMIM and separated mutations associated with autosomal dominant or autosomal recessive inheritance. A total of 1777 gene-to-OMIM disease entries, which involve 1281 genes, 1466 OMIM disease IDs and 35154 mutations, are associated with either autosomal dominant or autosomal recessive inheritance.

Fraction of ‘in-frame’ mutations

We grouped missense and small in-frame insertions, deletions and indels (types of mutations as defined in HGMD) as ‘in-frame’ mutations, whereas nonsense, splicing and small out-of-frame frame insertions, deletions and indels we grouped as ‘truncating’ mutations. We calculated the fraction of ‘in-frame’ mutations as the number of ‘in-frame’ mutations divided by the total number of mutations in each gene for each mode of inheritance (Figure 2C and D and Supplementary Figures S1 and S8) or for each disease (Figure 5B). To minimize the possibility of any existing trend being obscured by genes with few mutations, we limited our analysis to genes that have five or more mutations associated with each inheritance (Figure 2C and D and Supplementary Figures S1 and S8) or each disease (Figure 5B).

Essential human genes were estimated from the orthologs of mouse (Goh *et al*, 2007), fly, worm and yeast essential genes. Fly essential genes were extracted from Flybase (Wilson *et al*, 2008b; phenotype class: ‘lethal’), yeast essential genes from SGD (Ball *et al*, 2000; phenotype: ‘inviable’), and worm essential genes from RNAiDB (Gunsalus *et al*, 2004; phenotypes: ‘lethal’, ‘embryonic lethal’, ‘larval lethal’ and ‘adult lethal’).

Profiling interaction defects of mutant proteins

Disease mutant clones were generated by PCR mutagenesis essentially as described previously (Suzuki *et al*, 2005). Forward and reverse internal primers used are listed (Supplementary Table 4). All sequence-confirmed Entry clones of mutant alleles were transferred individually by Gateway recombinational cloning into both pDB-dest and pAD-dest-CYH destination vectors, generating DB–ORF allele and AD–ORF allele fusions (Rual *et al*, 2005). To test against wild-type interactors, the DB–ORF and AD–ORF clones for CBS, HGD, ACTG1, CDK4 and PRKAR1A mutant proteins were transformed into *MAT*α MaV203 or *MAT*a MaV103 yeast strains, respectively. Each interaction pair was tested for growth on SC-His + 3AT (synthetic medium without leucine, tryptophan and histidine, containing 20 mM 3-amino-1,2,4-triazole) plates to confirm *GAL1*::*HIS3* transcriptional activity, on yeast extract–peptone–dextrose (YPD) medium to determine *GAL1*::*lacZ* transcriptional activity using a -galactosidase filter assay, and on SC-Ura plates (synthetic medium without leucine, tryptophan and uracil) to determine *SPAL10*::*URA3* transcriptional activity. Scoring of Y2H reporters was done by comparing to a set of Y2H control strains that contain plasmids expressing pairs of proteins with a spectrum of interaction strengths (Supplementary Figure S9). Activation of at least two of the three reporter genes was taken as a positive interaction. Interaction pairs showing less than two positive reporters are scored as ‘−’. Interaction pairs showing the same number of positive reporters as the corresponding wild type are scored as ‘+’. Interactions that lose expression of one reporter but still show expression of the other two reporters are scored as ‘R’.

For immunoblotting, yeast cells with AD–ORF fusions were cultured overnight at 30°C in synthetic medium without tryptophan and then grown in YPD medium to mid-exponential phase. Cells were collected and treated with 150 mM of NaOH on ice for 15 min and then lysed in 0.8% SDS buffer (0.024 M Tris–HCl (pH 6.8), 10% glycerol, 0.04% bromophenol blue and 0.4% 2-mercaptoethanol) for 5 min at 95°C. Whole cell lysates were cleared by centrifugation at 14 000 *g*. Resulting supernatants were separated on NuPAGE acrylamide gels (Invitrogen) and electrophoretically transferred onto a PVDF membrane (Invitrogen). AD fusion proteins were detected by standard immunoblotting techniques using anti-GAL4 (Activation domain) antibody produced in rabbit (Sigma) as the primary antibody.

For comparison with experimental data, the following structures were used: 1JBQ for CBS (Meier *et al*, 2001), 1EYB and 1EY2 for HGD (Titus *et al*, 2000), 2BTF (Schutt *et al*, 1993), 1HLU (Chik *et al*, 1996) and 2OAN (Lassing *et al*, 2007) for bovine β-actin, 2W9F, 2W9Z, 2W96, 2W99 (Day *et al*, 2009) for CDK4, and 1G3N (Jeffrey *et al*, 2000) for CDK6–CDKN2C complex. Figures of tertiary structures were generated with PyMol (http://www.pymol.org). The relative solvent-accessible surface areas (%ASAs) were calculated with PSAIA (Mihel *et al*, 2008).

Structural analyses

Protein structures were downloaded from the Protein Data Bank website (PDB, http://www.rcsb.org/pdb). Removal of redundant structures was achieved using the PISCES server (Wang and Dunbrack, 2005) with the following criteria: X-ray structures only; no structure with Cα only; resolution ⩽3 Å; R-factor ⩽0.3; sequence length between 40 and 10 000 amino acids; and maximum 90% of sequence identity between similar PDB structures. This filtering collected 249 non-redundant protein structures corresponding to 236 genes in HGMD. To repair residual mismatches between the residue numeration in PDB files and in HGMD, PDB sequences were aligned against their corresponding cDNA sequences in HGMD using CLUSTALW (Chenna *et al*, 2003). The relative accessibility of over 91 000 residues in all 249 structures was calculated using PSAIA (Mihel *et al*, 2008). With multimers, accessibility was computed for all monomers considered independently and the multiple values obtained for the same residue were averaged. Among the 3664 residues affected by missense mutations, 1590 and 1045 were associated with autosomal recessive and autosomal dominant diseases, respectively.

Pfam domain assignment

Pfam domains (Pfam-A family only) were computed for cDNA sequences provided by HGMD, using InterProScan version 4.3 (http://www.ebi.ac.uk/Tools/InterProScan/). Missense, nonsense, in-frame and out-of-frame small insertions, deletions, and indels were then mapped onto the cDNA sequences and Pfam domains, generating a dataset containing 1348 genes with at least one Pfam-A domain and 34 964 associated mutations. Among them, a total of 10 904 'truncating' mutations are used for the analysis shown in Figure 4D, including 6212 associated with autosomal dominant diseases and 4692 associated with autosomal recessive diseases. Statistics were generated on the sum of a particular mutation type that either fell into or out of any Pfam-A domain in its respective protein versus the total fraction of the Pfam-A domain sequences in the protein sequence.

Transcription factors and structural proteins

Information on genes encoding transcription factors was obtained from Gene Ontology (Harris *et al*, 2004) annotations (948 genes with the GO term of 'transcription factor activity') and predictions in the transcription factor database (DNA Binding Domain, DBD; Wilson *et al*, 2008a; 1467 genes). A total of 1697 human transcription factor genes were retrieved. Among them, 82 genes associated with autosomal dominant diseases that have at least one mutation in HGMD were used for Pfam analysis (Figure 4D), and 56 genes with five mutations or more were used for analysis of 'in-frame' mutations (Figure 2D). Structural protein coding genes were retrieved from Gene Ontology annotations of 'cytoskeleton' (992 genes). Among them, 72 genes with at least one mutation in HGMD were used for Pfam analysis (Figure 4D), and 47 genes with five mutations or more were used for analysis of 'in-frame' mutations (Figure 2D). DBD and Gene Ontology data were downloaded in March 2008.

Statistical analysis

Error bars represent the s.e.m. values. Significance of the observed difference in the distributions of 'in-frame' versus 'truncating' mutations in autosomal dominant and autosomal recessive disease, the greater proportions of 'in-frame' mutations in structural proteins than in transcription factors, as well as the greater accessibility of residues mutated in autosomal dominant versus autosomal recessive diseases, was evaluated using the non-parametric Mann–Whitney *U* test. Enrichments of disease alleles in Pfam domains were determined using odds ratio and the significance thereof using Fisher's exact test. A fold enrichment higher than one means Pfam domains contain more mutations than expected at random, whereas an enrichment between zero and one means a depletion in mutations. The differences between proportions of 'in-frame' mutations in each pair of diseases associated with the same gene were assessed by Fisher's exact test. All statistics were computed using the R package (http://www.r-project.org/).

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank all members of the Vidal Lab and the Center for Cancer Systems Biology (CCSB), Dr Patricia K Donahoe and Dr Roseann Mulloy for helpful suggestions; Ines M Pinto for help with experiments. This study was supported by the Ellison Foundation and the WM Keck Foundation (MV), NIH grants R01-HG001715 from NHGRI (MV and F Roth), U01-CA105423 (PI, S Orkin, project leader, MV), U54-CA112952 (PI, J Nevins, subcontract, MV) and R33-CA132073 (MV) from NCI, and by Institute Sponsored Research funds from the Dana-Farber Cancer Institute Strategic Initiative awarded to CCSB. KV was supported by an NIH NRSA training grant fellowship (T32-CA09361). BC was supported by the Belgian Program on Interuniversity Attraction Poles initiated by the Federal Office for Scientific, Technical and Cultural Affairs (IAP P6/19 PROFUSA). MV and RB are 'Honorary Research Associate' and 'Research Director' from the Fonds de la Recherche Scientifique (FRS-FNRS, French Community of Belgium), respectively.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Ball CA, Dolinski K, Dwight SS, Harris MA, Issel-Tarver L, Kasarskis A, Scafe CR, Sherlock G, Binkley G, Jin H, Kaloper M, Orr SD, Schroeder M, Weng S, Zhu Y, Botstein D, Cherry JM (2000) Integrating functional genomic information into the *Saccharomyces* genome database. *Nucleic Acids Res* **28:** 77–80

Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **33 :** 228–237

Braun P, Tasan M, Dreze M, Barrios-Rodiles M, Lemmens I, Yu H, Sahalie JM, Murray RR, Roncari L, de Smet AS, Venkatesan K, Rual JF, Vandenhaute J, Cusick ME, Pawson T, Hill DE, Tavernier J, Wrana JL, Roth FP, Vidal M (2009) An experimentally derived confidence score for binary protein–protein interactions. *Nat Methods* **6:** 91–97

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* **31:** 3497–3500

Chik JK, Lindberg U, Schutt CE (1996) The structure of an open state of β-actin at 2.65 Å resolution. *J Mol Biol* **263:** 607–623

Day PJ, Cleasby A, Tickle IJ, O'Reilly M, Coyle JE, Holding FP, McMenamin RL, Yon J, Chopra R, Lengauer C, Jhoti H (2009) Crystal structure of human CDK4 in complex with a D-type cyclin. *Proc Natl Acad Sci USA* **106:** 4166–4170

Dreze M, Charloteaux B, Milstein S, Vidalain PO, Yildirim MA, Zhong Q, Svrzikapa N, Romero V, Laloux G, Brasseur R, Vandenhaute J, Boxem M, Cusick ME, Hill DE, Vidal M (in press) 'Edgetic' perturbation of a *C. elegans* BCL2 ortholog. *Nat Methods*

Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A (2006) Pfam: clans, web tools and services. *Nucleic Acids Res* **34:** D247–D251

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL (2007) The human disease network. *Proc Natl Acad Sci USA* **104:** 8685–8690

Gunsalus KC, Yueh WC, MacMenamin P, Piano F (2004) RNAiDB and PhenoBlast: web tools for genome-wide phenotypic mapping projects. *Nucleic Acids Res* **32:** D406–D410

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33:** D514–D517

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M *et al* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32:** D258–D261

Hsu WT, Pang CN, Sheetal J, Wilkins MR (2007) Protein-protein interactions and disease: use of *S. cerevisiae* as a model system.. *Biochim Biophys Acta* **1774:** 838–847

Jeffrey PD, Tong L, Pavletich NP (2000) Structural basis of inhibition of CDK–cyclin complexes by INK4 inhibitors. *Genes Dev* **14:** 3115–3125

Lamesch P, Li N, Milstein S, Fan C, Hao T, Szabo G, Hu Z, Venkatesan K, Bethel G, Martin P, Rogers J, Lawlor S, McLaren S, Dricot A, Borick H, Cusick ME, Vandenhaute J, Dunham I, Hill DE, Vidal M (2007) hORFeome v3.1: a resource of human open reading frames representing over 10 000 human genes. *Genomics* **89:** 307–315

Lassing I, Schmitzberger F, Bjornstedt M, Holmgren A, Nordlund P, Schutt CE, Lindberg U (2007) Molecular and structural basis for redox regulation of β-actin. *J Mol Biol* **370:** 331–348

Maglott D, Ostell J, Pruitt KD, Tatusova T (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* **33:** D54–D58

Meier M, Janosik M, Kery V, Kraus JP, Burkhard P (2001) Structure of human cystathionine β-synthase: a unique pyridoxal 5′-phosphate-dependent heme protein. *EMBO J* **20:** 3910–3916

Mihel J, Sikic M, Tomic S, Jeren B, Vlahovicek K (2008) PSAIA—protein structure and interaction analyzer. *BMC Struct Biol* **8:** 21

Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S *et al* (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437:** 1173–1178

Sammut SJ, Finn RD, Bateman A (2008) Pfam 10 years on: 10 000 families and still growing. *Brief Bioinform* **9:** 210–219

Schuster-Böckler B, Bateman A (2008) Protein interactions in human genetic diseases. *Genome Biol* **9:** R9

Schutt CE, Myslik JC, Rozycki MD, Goonesekere NC, Lindberg U (1993) The structure of crystalline profilin–β-actin. *Nature* **365:** 810–816

Seidman JG, Seidman C (2002) Transcription factor haploinsufficiency: when half a loaf is not enough. *J Clin Invest* **109:** 451–455

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* **21:** 577–581

Subramanian S, Kumar S (2006) Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. *BMC Genomics* **7:** 306

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

Titus GP, Mueller HA, Burgner J, Rodriguez De Cordoba S, Penalva MA, Timm DE (2000) Crystal structure of human homogentisate dioxygenase. *Nat Struct Biol* **7:** 542–546

van Bokhoven H, Brunner HG (2002) Splitting p63. *Am J Hum Genet* **71:** 1–13

Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh KI, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie JM, Cevik S, Simon C, de Smet AS, Dann E, Smolyar A *et al* (2009) An empirical framework for binary interactome mapping. *Nat Methods* **6:** 83–90

Wang G, Dunbrack Jr RL (2005) PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res* **33:** W94–W98

Wang Z, Moult J (2001) SNPs, protein structure, and disease. *Hum Mutat* **17:** 263–270

Wilkie AO (1994) The molecular basis of genetic dominance. *J Med Genet* **31:** 89–98

Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA (2008a) DBD—taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res* **36:** D88–D92

Wilson RJ, Goodman JL, Strelets VB (2008b) FlyBase: integration and improvements to query tools. *Nucleic Acids Res* **36:** D588–D593

Ye Y, Li Z, Godzik A (2006) Modeling and analyzing three-dimensional structures of human disease proteins. *Pac Symp Biocomput* **2006:** 439–450

Yue P, Li Z, Moult J (2005) Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol* **353:** 459–473