Network properties of genes harboring inherited disease mutations

Igor Feldman*, Andrey Rzhetsky*†‡, and Dennis Vitkup*‡

*Department of Biomedical Informatics, Center of Computational Biology and Bioinformatics, Columbia University, New York, NY 10032; and †Computation Institute and Institute of Genomics and Systems Biology, University of Chicago, Chicago, IL 60637

Edited by Randy Schekman, University of California, Berkeley, CA, and approved January 17, 2008 (received for review February 23, 2007)

By analyzing, in parallel, large literature-derived and high-throughput experimental datasets we investigate genes harboring human inherited disease mutations in the context of molecular interaction networks. Our results demonstrate that network properties influence the likelihood and phenotypic consequences of disease mutations. Genes with intermediate connectivities have the highest probability of harboring germ-line disease mutations, suggesting that disease genes tend to occupy an intermediate niche in terms of their physiological and cellular importance. Our analysis of tissue expression profiles supports this view. We show that disease mutations are less likely to occur in essential genes compared with all human genes. Disease genes display significant functional clustering in the analyzed molecular network. For about one-third of known disorders with two or more associated genes we find physical clusters of genes with the same phenotype. These clusters are likely to represent disorder-specific functional modules and suggest a framework for identifying yet-undiscovered disease genes.

computational biology | disease genes | systems biology

The impact of a single-nucleotide substitution can be markedly different depending on where it occurs in the human genome. The substitution may lead to no detectable effect if, for example, it falls within a noncoding sequence or a third codon position of a gene. When a harmful substitution does affect protein or RNA function, a continuum of outcomes ranging in phenotypic severity is possible. In the worst case, the nucleotide change is lethal (the corresponding gene is often classified as essential), and the organism dies early in its development. A milder but still observable phenotypic effect is what we usually call a disease (the corresponding gene is classified as a disease gene): a significant nonlethal malfunction within the human physiological system. Yet a milder physiological effect may be invisible in all but rare situations, for example, the inability to recognize a specific odor. A similar spectrum exists for the favorable genetic changes, but they are likely to be rarer and are less studied. Clearly, the strength of phenotypic consequences of a genetic variation is continuous; the three classes of phenotypic outcomes outlined above are but products of a somewhat arbitrary partition of this natural continuum. It is also true that different mutations within the same gene (for example, *p53*) can lead to outcomes spanning the entire phenotypic spectrum.

It is likely that a gene location within a cellular network may influence the impact and consequences of a given gene mutation. Here, we test this hypothesis by analyzing a large collection of known human disease genes in the context of several human molecular networks.

Human Interactome Data

Currently available human molecular interaction networks are neither complete, nor error-free. Nevertheless, several recent studies generated large datasets approximating the complete human interactome. In our analysis we used three human protein interaction datasets. One is a large-scale dataset of physical interactions extracted from hundreds of thousands of full-length scientific articles by the GeneWays (GW) natural language system (1, 2). The GW interaction network contains 4,458 human genes and 12,991 physical interactions (such as phosphorylate or bind) between them. The other two interaction networks were generated by two experimental studies using the yeast two-hybrid (Y2H) technique, one by Stelzl *et al.* (3) with 1,693 genes and 3,120 interactions, and the other by Rual *et al.* (4) with 1,549 genes and 2,611 interactions. To improve the statistical power in our analysis we combine the two Y2H datasets into a joint Y2H network with 2,965 nodes and 5,722 interactions.

We view the Y2H and GW networks as complementary rather than competing views of the human interactome, much like two photographs of the same landscape from different viewpoints. There are likely biases in both types of the networks: for example, interactions between membrane-bound proteins tend to be underdetected with two-hybrid screens, whereas the literature-derived interactions are likely to overrepresent interactions between well studied proteins (3–9). Although we observe similar trends for both GW and Y2H datasets, the smaller size of the Y2H network leads to less statistically significant results compared with the GW network. Because the total number and types of interactions discovered by the two-hybrid methods and literature mining are different, the Y2H and GW networks are not directly comparable to one another (e.g., in terms of the average network connectivity).

Phenotype Data

In our work we analyze a large compendium of genes harboring known inherited disease mutations compiled by Jimenez-Sanchez *et al.* (10). The set contains 908 disease genes, of which 498 and 144 can be mapped to GW and Y2H networks, respectively. The vast majority of disease genes in the set are responsible primarily for monogenic (Mendelian) disorders. Nevertheless, we could clearly identify 38 and 20 genes associated with polygenic disorders in the GW and Y2H networks, respectively. The majority of the polygenic disease genes (76% on GW and 65% on Y2H) are associated with various forms of cancer. The disease gene set that we analyze is not complete, and future studies will undoubtedly identify additional disease genes. Therefore, it is appropriate to view our analysis as an attempt to estimate from the incomplete data the probability that a random mutation in a gene with certain network properties would lead to an inherited disease (as opposed to a lethal or nondisease phenotype). For comparison, we also analyzed the human orthologs of essential mouse genes (806 and 298 genes mapped to the GW and Y2H networks, respectively) from the Mouse Genome Database (11). We use these orthologs here as an approximation for the set of essential human genes. The sets of disease and essential genes are

Author contributions: A.R. and D.V. designed research; I.F., A.R., and D.V. performed research; I.F., A.R., and D.V. analyzed data; and I.F., A.R., and D.V. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

‡To whom correspondence may be addressed. E-mail: arzhetsky@uchicago.edu or dennis.vitkup@dbmi.columbia.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0701722105/DC1.

www.pnas.org/cgi/doi/10.1073/pnas.0701722105

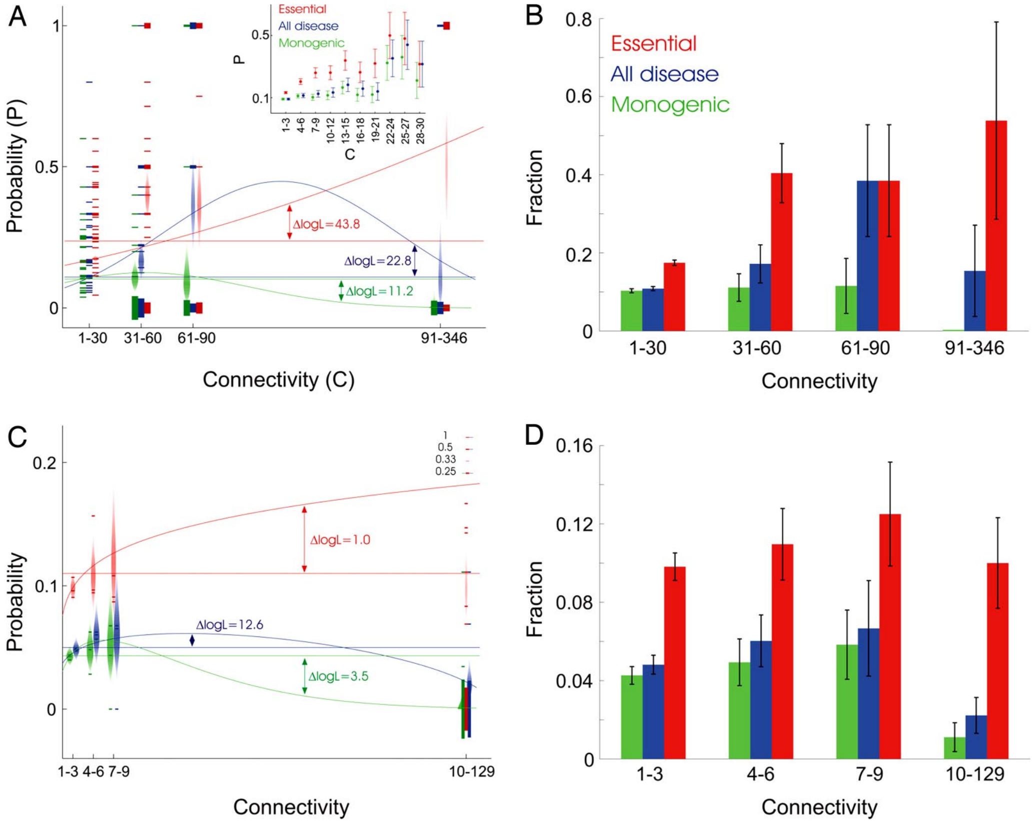

Fig. 1. The probabilities (fractions) for three gene categories: monogenic disease (green), all disease (blue), and essential (red) as a function of their network connectivity. (*A*) Fit of the probabilistic models to the GW network data. The curves represent the maximum-likelihood model fits to all (nonbinned) data (see *Methods*). Separately for monogenic and all disease genes, two models describing the data were tested: a general model using a bell β-like function and a uniform null hypothesis model. For essential genes, a rising β-like function was tested against a uniform null hypothesis. The log-likelihood differences between the models are shown next to the corresponding arrows. The individual data points representing the fractions of all disease, monogenic disease, and essential genes at each connectivity are shown by green, blue, and red dots. The data points were collected to four bins for display purposes only. For each bin, 99% confidence intervals for the posterior probabilities are represented by colored densities. The color intensity and width of the densities represent the probability values. (*Inset*) The data for the first bin. The error bars represent SEM. (*B*) A simpler bar plot of the GW network data presented in *A*. The fractions of all disease genes, monogenic disease genes, and essential genes are shown for different gene connectivity bins. The error bars represent SEM. (*C*) The same as *A* for the Y2H network with fit of the probabilistic models to the Y2H network data. (*D*) A simpler bar plot of the Y2H network data presented in *C*. The fractions of all disease genes, monogenic disease genes, and essential genes are shown for different gene connectivity bins. The error bars represent SEM.

not mutually exclusive. Some of the disease genes under a complete knockout produce a lethal phenotype and can be as well characterized as essential. Whether a gene is essential or a disease one is determined by a particular mutation.

Results and Discussion

Network Properties of Disease Genes. To examine topological network properties of disease genes we first use a number of established network statistics, such as node connectivity (degree) and clustering coefficient. A node's connectivity is defined as the total number of edges connecting it with other network nodes (12). We find that connectivity of polygenic disease genes is significantly higher than that of other disease genes. The GW (Y2H) average polygenic and monogenic disease gene connectivities are 38.3 and 6.8 (7.7 and 2.7), respectively (Mann–Whitney's $P < 0.0001$ and 0.08, respectively). The observed difference primarily comes from the highly connected signaling proteins that are responsible for various forms of cancer.

The average connectivity of disease genes is significantly higher than that of an average gene in the GW network (disease 9.3, all 5.9, $P < 10^{-10}$) and about equal for the Y2H network (disease 3.4, all 3.9, $P = 0.28$). This difference in the average disease gene connectivities for the two networks could be a consequence of the knowledge bias (disease genes are, on average, better studied) present in the GW network. To investigate further the connectivity distribution of disease genes, we calculate the probability to find a disease gene occupying a network node with a given connectivity (see Fig. 1 *A* and *C*). Using the same data we also show the binned distributions, separately for the GW and Y2H networks, in Fig. 1 *B* and *D*. For comparison, we plot the probability (fraction for the bar graphs) of essential genes in the figures. For the lower half of the connectivity bandwidth in Fig. 1, the mean fractions of both disease and essential genes increase with the number of interactions for both the GW and Y2H networks (see also Fig. 1*A Inset*). However, for higher connectivities, we can see that the fraction of essential genes is still high, whereas that of disease genes drops significantly. We observe a similar behavior for the monogenic genes, although the rise for low connectivities is not pronounced for the Y2H network. The relationship between the network connectivity and gene essentiality has been discussed previously. Highly

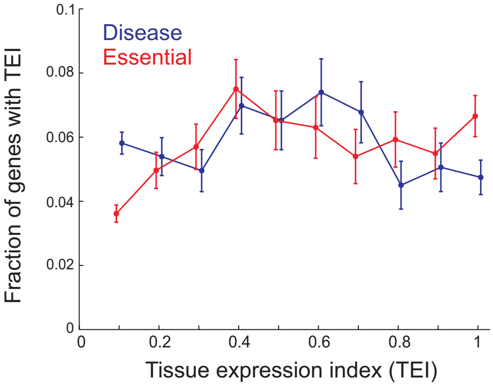

Fig. 2. Tissue expression distribution for disease and essential genes. The tissue expression index (TEI) was calculated for every gene as the fraction of the 79 tissues analyzed by Su *et al.* (18) in which the gene was detected as expressed. The genes with large indexes are expressed in almost all tissues, whereas the genes with small indexes have a limited expression distribution. Shown are the fractions for disease and essential genes as the function of tissue expression index. The error bars represent SEM.

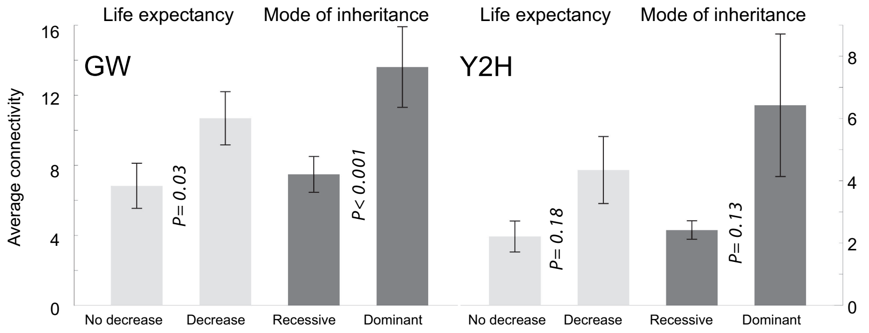

Fig. 3. Average network connectivity for disease genes with different phenotypes. Data are shown separately for the GW (*Left*) and Y2H (*Right*) networks. Light-gray columns show the average connectivity for disease genes displaying decrease/no-decrease in life expectancy. Dark-gray columns show the average connectivity of disease genes with recessive/dominant phenotypes. The error bars represent one standard error. Because the connectivity distributions for each category are not parametric, we used the Mann–Whitney test to determine significance of the difference between categories.

connected proteins were found to be, on average, more essential (13), although the relationship between a gene's essentiality and connectivity is not deterministic or simple (5, 14). The observation that nonsomatic disease genes are less likely to be observed in networks' hubs (highly connected nodes) suggests the less damaging nature of disease mutations, likely allowing the organism's survival.

A natural way to test the robustness of the apparent bell-shaped probability distribution of disease genes is to design a probabilistic model incorporating multiple distribution shapes. We can then rigorously compare different distribution shapes in terms of the fit quality to the real data (using the likelihood ratio test). Therefore, we implement two nested parametric models for the observed data: the simpler model describes a uniform distribution of disease genes across the connectivity range, whereas the more general one allows for the distribution to be uniform-like, bell-shaped, or rising (a beta distribution-like function; see *Methods*). Applying the models to real data, we find that the likelihood of the more general bell-shaped model is much greater than that of the uniform model for disease genes (P values reflecting the significance of the difference are 10^{-5} and 0.002 for the GW and Y2H networks, respectively). We obtain similar results by using a simpler model [see supporting information (SI) *Appendix*] with fewer parameters in which high statistical significance was observed in all but one case (disease genes on Y2H network; $P = 0.2$). It is very likely that the disease genes more often get in the limelight of scientific studies than other genes (for example, the focus on disease genes may influence the subsequent funding of the research project). Although this attention bias may explain a higher-than-average connectivity of disease genes in the GW network, it clearly cannot explain the low probability of finding disease genes at the high-connectivity part of the distribution (network hubs). In line with previous observations (13, 15), the likelihood of the rising model is higher than that of the uniform one for the essential genes ($P = 4 \times 10^{-11}$ and 0.3 for the GW and Y2H networks, respectively). To investigate the possible biases associated with selection of candidate disease genes through functional genomics methods (16), we repeated the analysis by using disease genes identified through positional cloning and the Y2H network (see SI Table 1 and *SI Appendix*); the resulting distribution is similar to the one obtained for all disease genes (a bell-shaped curve fit is significantly better than a uniform one ($P = 0.05$).

Important cellular complexes and functional modules often correspond to dense regions in protein interaction networks, i.e., nodes with high connectivities and high clustering coefficients. Clustering coefficient of a node is defined as the ratio between the observed number of direct connections between the node's immediate network neighbors to the maximum possible number of such connections (12). A higher clustering coefficient for a node corresponds to a higher density of network connections around it. Computing clustering coefficients for both GW and Y2H networks, we observe that disease genes "avoid" dense-clustering neighborhoods unlike the essential genes. The average clustering coefficient for highly connected nodes (i.e., nodes likely to represent functional modules and complexes) for the GW network were 0.11 for disease and 0.14 for essential genes ($P = 0.005$) and for the smaller and sparser Y2H network they were 0.015 for disease and 0.030 for essential genes ($P = 0.7$). The nodes with connectivity >6 were used to represent highly connected network clusters; similar results were obtained by using other connectivity thresholds (see *SI Appendix*).

Tissue Expression of Disease Genes. The tendency to escape most vital cellular components while affecting lesser physiological processes appears to be a general property of disease genes. If so, this property is likely to be observed with completely different descriptors of protein function, such as multitissue expression measurements. It has been shown previously that disease genes have a significantly narrower range of tissue expression compared with all genes (17). Here, we analyze the tissue distribution of expression for disease and essential genes (see Fig. 2), using data from Su *et al.* (18) for 79 human tissues. For each gene we calculate the fraction of human tissues in which it was significantly expressed. Similar to the network connectivity, genes with intermediate values of the expres-

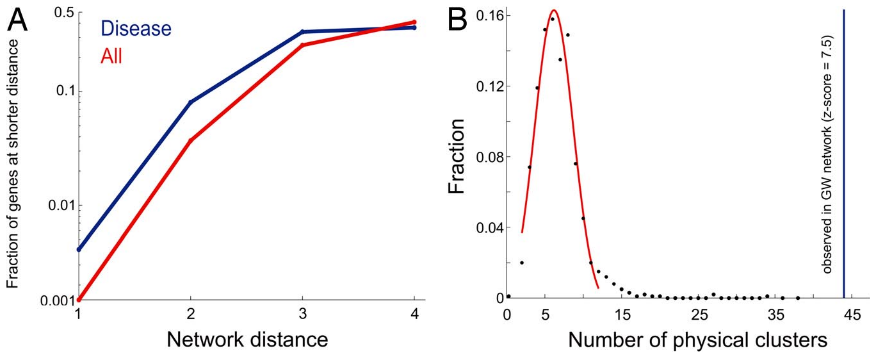

Fig. 4. Physical clustering of disease genes in the GW network. (*A*) The red line shows the fraction of all genes located at a certain network distance or closer. The shortest path between each gene pair was used to calculate the network distance. The blue line shows the fraction of all disease genes located at a certain network distance or closer. (*B*) Disorder-specific clustering of disease genes. In the GW network, there are 38 clusters of physically interacting genes associated with the same disorder. To investigate the significance of the observed clustering we simulated the distribution of the physical clusters between genes associated with the same disorder. The random distribution of clusters was obtained by reshuffling network edges while preserving the total connectivity of each gene. The reshuffling was repeated 1,000 times to obtain the distribution. The results of the network randomization show that the observed clustering is highly statistically significant (z score > 7.5). The Gaussian fit to the simulated distribution is shown with a solid red line.

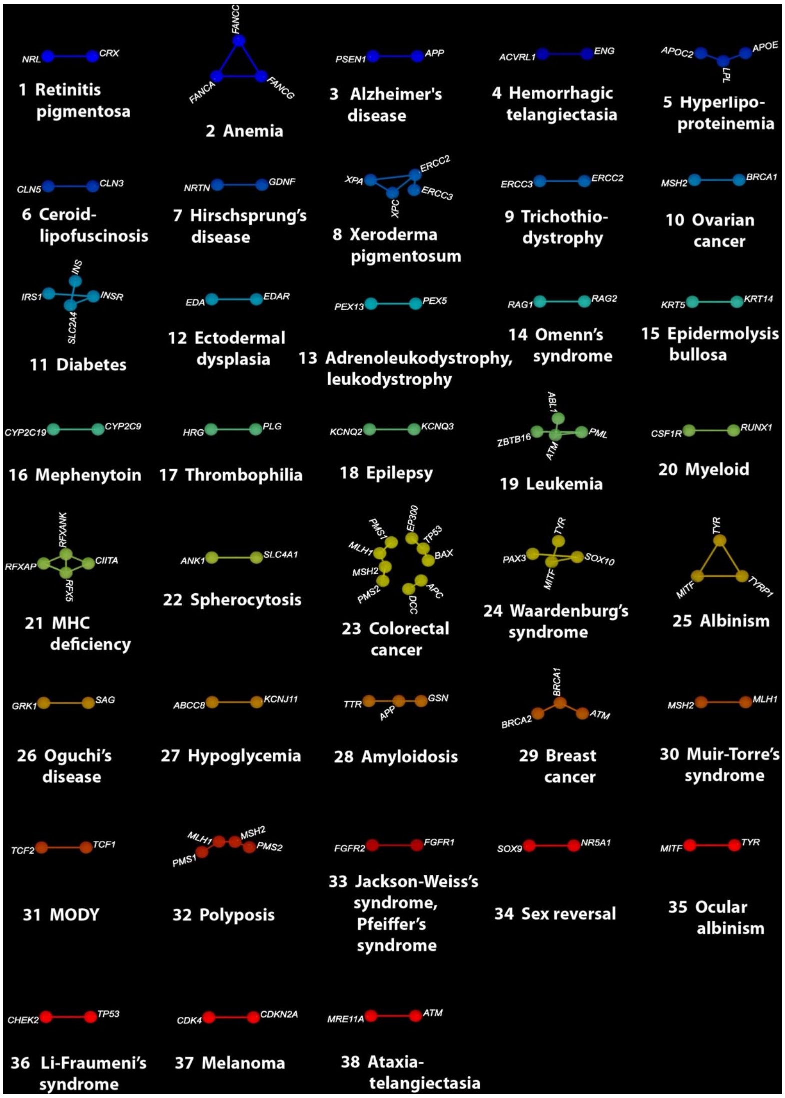

Fig. 5. Genes and proteins harboring variation causing the same disease phenotype tend to form directly (physically) connected clusters. Physical-interaction gene clusters associated with 38 disease phenotypes are shown. Gene and phenotype names are indicated for each cluster. The phenotype numbers and cluster colors serve as the key for Fig. 6.

sion fraction have the highest probability to harbor inherited disease mutations. The probability of disease genes to occupy the intermediate (0.4–0.7) range of the expression fractions is significantly higher compared with genes with the smaller (0–0.3; $P < 0.001$) or larger (0.8–1; $P < 0.001$) ranges. As with connectivity distributions, to make sure that the tissue expression results are not caused by the ascertainment biases, we repeat the same analysis with only positionally cloned disease genes. The resulting distributions show a similar pattern with a statistically significant peak at the intermediate expression range (see *SI Appendix*). It is likely that genes expressed in a small number of tissues are, on average, less physiologically important, whereas genes expressed in almost all tissues are on average too important to allow survival when damaged. Essential and disease genes again show qualitatively different behavior. The probabilities of essential genes to have intermediate and high expression levels are not significantly different from each other (P value close to 1) and are both significantly higher than the probabilities for essential genes to have a low expression ($P < 0.001$). The expression analysis reinforces the view that the inherited disease-causing mutations have the highest probability of occurring in genes with intermediate physiological importance.

Disease Versus Essential Genes. We observe different behaviors for disease and essential genes in terms of the connectivity and expression distributions. In this context it is interesting to investigate whether disease mutations preferentially occur in nonessential genes. Following this question, we compare the available sets of disease and essential genes directly by using the χ^2 test. The two sets are significantly different when using (as we do throughout this study) the human orthologs of mouse essential genes as the set of essential human genes ($P < 10^{-12}$). To check this result for possible ascertainment biases in the selection of mouse essential genes, we also repeat the test by using human orthologs of essential *Caenorhabditis elegans* genes identified in a systematic all-genome RNAi screen (19). We again observe that the two sets are highly significantly different ($P = 10^{-4}$) (see *SI Appendix* for details). Consequently, disease mutations do have a lower probability of occurring in essential genes.

Severity of Mutation Outcome and Network Topology. It was demonstrated previously that the molecular function of a disease gene affects its mode of inheritance (20). If the network connectivity affects the likelihood of a mutation to cause a disease phenotype,

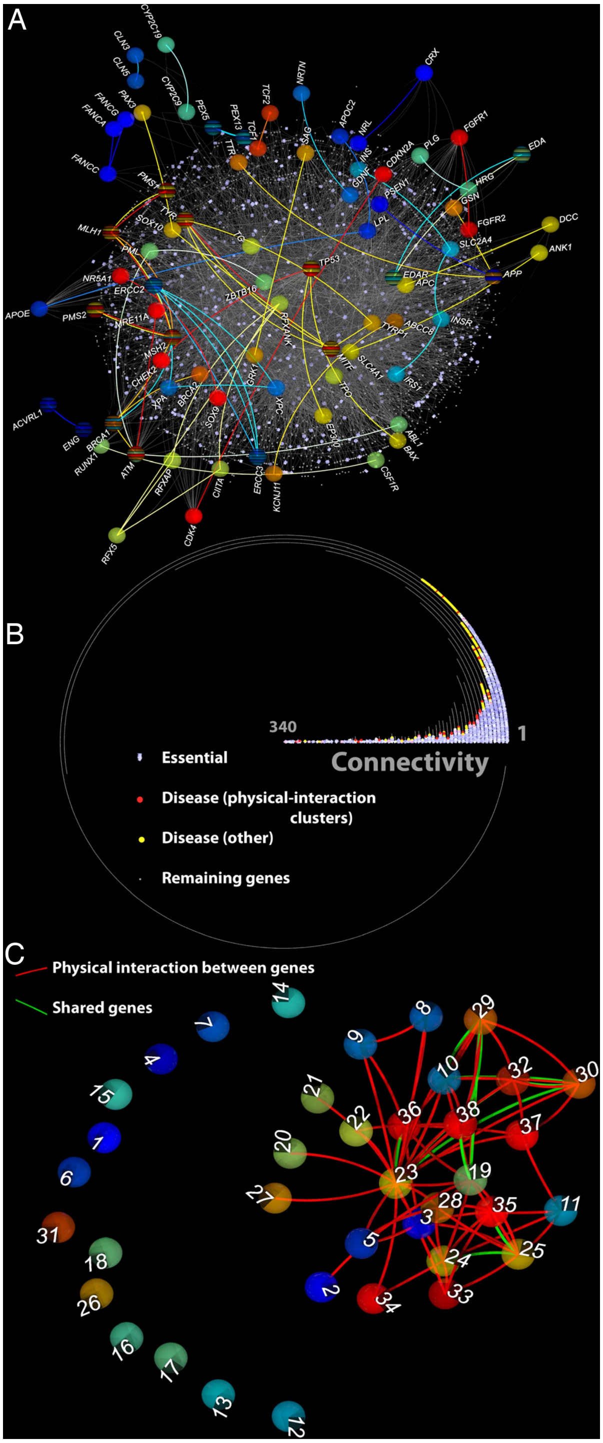

Fig. 6. Genes and proteins harboring variation causing the same disease phenotype tend to form directly (physically) connected clusters (continued from Fig. 5). (*A*) A visualization of the same 38 phenotypic gene clusters as shown in Fig. 5 within the GW molecular interaction network. Genes associated with the same phenotype are indicated by the same-color semitransparent spheres. Note that several genes within the network are known to affect multiple phenotypes (network nodes with multicolor stripes). The blue cubes represent essential genes and provide additional network context. (*B*) A detailed view of gene connectivity distribution (compare with Fig. 1 *A* and *B*) in the GW network. Disease (red and yellow spheres), essential (blue cubes), and other (white dots) genes are placed along concentric circles that represent gene connectivity layers within the molecular network. In our GW network the connectivity covers range between 1 (the outermost circle) and 340 (the center). We can see that the intermediate connectivity range contains a higher proportion of disease genes participating in physical clustering (red spheres) than the disease genes that do not (yellow spheres). Subplots (*A* and *C*) focus exclusively on the subset of disease genes that have within-phenotype physical-interaction clustering. (*C*) The overlaps and physical interactions between gene clusters linked to 38 disease phenotypes. Each node represents a whole disease cluster of the same color and number as used in Fig. 5. Two nodes are connected by a red edge when there is at least one direct physical interaction that links two genes from the two distinct phenotypic clusters represented by the graph nodes. Two nodes are connected by a green edge when their corresponding disease clusters share at least one gene.

it may also influence the severity of the disease outcome. To test this hypothesis we analyze the phenotypic characteristics of disease mutations compiled by Jimenez-Sanchez *et al.* (10). Our results (see Fig. 3) reveal that genes harboring disease mutations with dominant phenotype display significantly higher network connectivity compared with genes with recessive phenotype ($P < 0.001$ for GW; $P = 0.13$ for Y2H). This result may be a consequence of larger network perturbations introduced by mutations in highly connected genes. Moreover, we also observe that mutations reducing life expectancy reside in more connected genes compared with mutations without such an effect ($P = 0.03$ for GW; $P = 0.18$ for Y2H). Larger decrease in life expectancy caused by mutations in highly connected genes again suggests that these mutations are, on average, more damaging and lead to larger physiological consequences, than mutations in genes with lower connectivity.

Physical Clustering of Disease Genes. Our work would be incomplete without an analysis of the relative network distribution of disease genes. Do they have a higher probability to interact with one another, or they are spread evenly around the network? Do they form large connected clusters? Are these clusters distributed homogeneously across the network? To address these questions, we focus on the GW network because it includes 460 direct interactions between disease genes (versus only 9 in the significantly less dense Y2H network). As an instrument for the analysis of interaction patterns of disease genes we design two nested probabilistic models. In the first model disease genes are blind to the type (disease or nondisease) of the interacting neighbor. In the second model disease genes preferentially interact with each other; their self-affinity is described by parameter θ (see *SI Appendix*). We find that the preferential affinity model describes the data with significantly higher likelihood ($P < 0.001$) and suggests that disease genes are about two times more likely to interact with each other as compared with nondisease genes ($\theta = 0.20$ vs. 0.11). A consequence of this preferential interaction is a skewed distribution of the shortest network distances between disease genes (see Fig. 4*A*): 8.0% of the disease genes are located within network distance 2, compared with average 3.7% for the whole network.

In the GW network disease genes form a large connected cluster of 271 nodes and several smaller clusters (2 of size 3 and 14 of size 2), and 192 genes are not connected to other disease genes. The observed large cluster does not group genes functionally (because it contains genes from many unrelated diseases), but indicates that complex phenotypes are entangled genetically. The formation of such a cluster could be a consequence of the existence of a giant connected component in a densely connected network (12).

We are particularly interested in physical-interaction clustering of same-phenotype genes: our dataset contains 128 groups of two or more genes associated with the same phenotype (such as breast cancer, Waardenburg's syndrome, and retinitis pigmentosa; see Figs. 5 and 6 and *SI Appendix*). This disease-focused clustering is different from gene clustering into broader physiological categories, such as developmental, nervous system, and metabolic, demonstrated by Gandhi *et al.* (5). We observe 38 connected clusters (of size 2 and larger) of genes associated with the same disorder (see Fig. 5). We evaluate the statistical significance of such clustering by randomly shuffling protein interactions in the GW network while maintaining the total connectivity of each protein. The results of such randomization (see Fig. 4*B*) demonstrate that the observed clustering is highly statistically significant (z score > 7.5; $P < 3.2 \times 10^{-14}$).

Importantly, about one-third of all studied disorders (38 of 128) with two or more associated genes show significant clustering in the interaction network. Of all phenotypes that we consider here, 8 of 30 (27%) of non-Mendelian and 30 of 99 (30%) of Mendelian ones map to sets of genes that form physical clusters. On the gene level, 22% (108 of 497) of all disease genes cluster with other genes responsible for the same disorder. In Figs. 5 and 6 we show disorder-specific gene clusters mapped onto the GW network.

In Fig. 5 we display all disorders with multiple associated genes forming physical connections (38 clusters). Fig. 6*A* shows the connections between multiple genes associated with the same disorder in the context of the GW network. Several genes within the network affect multiple phenotypes (shown with multicolor stripes in Fig. 6*A*), and genes responsible for different phenotypes often form connections. To highlight the connectivity distribution of genes participating in physical clusters we show them separately (Fig. 6*B*, red) from the rest of the disease gene (Fig. 6*B*, yellow). The genes with intermediate connectivity range contain a higher fraction of the clustering genes. To explore the connections between various phenotypes with multiple associated genes we show dependencies between phenotypes in Fig. 6*C*. In Fig. 6*C* each sphere represents one phenotypic cluster and the connections between phenotypes represent either shared genes (green edges) or physical interaction between genes from different phenotypes (red edges). Curiously, in the GW network 25 of 38 multigene phenotypic clusters form connections to other clusters.

If current estimates of the total size of the human interactome (21, 22) are correct, we analyzed at best 5–10% of all possible interactions between human proteins. It is therefore quite remarkable that we observe a high level of disease gene clustering despite a very limited knowledge of the human interactome. The observed physical clustering reflects functional similarity between genes associated with the same diseases; only about one-third of the functional clusters represent molecular complexes. As our knowledge of the human interactome expands, it is likely that the functional clusters responsible for specific diseases will grow and new clusters will emerge. The clustering of the disease genes in the interaction network is likely to be complemented by clustering of other functional categories, such as expression and subcellular localization. Analysis of such disease clusters and nearby genes is critically important as a guiding framework for identifying new disease genes (23, 24).

Methods

To compare the alternative interpretations of data, we designed three nested probabilistic models describing stochastic correspondence between the gene connectivity, c, and the number of (disease or essential) genes, d_c, with given connectivity. Let N_c be the total number of genes with connectivity c. In our description both d_c and d_c/N_c are random variables that follow different distributions for different values of c.

We assume that the mean value of d_c/N_c in each connectivity bin, $y_c = E(d_c/N_c|c)$, is deterministically defined by a β-function with parameters a, b, k, and ψ:

$$y_c = E\left[\frac{d_c}{N_c}\middle| C = c, a, b, k, \psi\right] = kc^{(a-1)}(\psi - c)^{(b-1)}. \qquad [1]$$

Further, we assume in all three models that given the value of y in the cth connectivity bin, y_c, the observed value of d_c follows a scaled β-distribution:

$$P(D_c = d_c|N_c, \alpha_c, \beta_c) = \binom{N_c}{d_c}\frac{B(\alpha_c + d_c, \beta_c + N_c - d_c)}{B(\alpha_c, \beta_c)}. \qquad [2]$$

where $B(x, y)$ represents a two-parameter β-function, where of two parameters α_c and β_c only one is free because of the constraint on the mean value of d_c, $E((d_c/N_c)|c) = y_c$:

$$\begin{aligned} \alpha_c &= \frac{(1 - y_c)y_c^2}{\sigma_c} - y_c, \\ \beta_c &= \frac{1 - y_c}{y_c}\alpha_c, \end{aligned} \qquad [3]$$

where $[\sigma_c]^2$ represents the variance of y_c.

$$\begin{aligned} \log L &= \sum_c \log[P(d_c|c, \Theta)], \\ \Theta_{ML} &= \underset{\Theta}{\arg\max}\,[\log L]. \end{aligned} \qquad [4]$$

We can obtain three nested models in the following way. For the general model, a and b are unconstrained. For the rising curve, $b = 1$. For uniform distribution, $a = b = 1$. Clearly, the rising curve model is nested to the general model, whereas the uniform distribution is nested to both more parameter-rich models. Then we can use the standard theory of asymptotic behavior of the maximum-likelihood values to assess the significance of difference in quality of fit to data of the alternative models.

Note Added in Proof. After this article was submitted to PNAS, we learned about a similar study by Goh *et al*. (25) that was submitted to PNAS at the same time as our work.

ACKNOWLEDGMENTS. We thank Murat Cokol, Michiel de Hoon, Ivan Iossifov, Raul Rodriguez-Esteban, Tzu-Lin Hsiao, and Chani Weinreb for comments on the manuscript; Murat Cokol and Raul Rodriguez-Esteban for comments on figure design; Tzu-Lin Hsiao for help with expression data; Ivan Iossifov for comments on mathematical modeling; and Gerardo Jimenez-Sanchez (National Institute of Genomic Medicine, Mexico DF, Mexico) for providing a set of positionally cloned genes. This work was supported by the National Institutes of Health Grant GM079759 (to D.V.), National Institutes of Health Grant GM61372 (to A.R.), National Science Foundation Grants 0438291 and 0121687 (to A.R.), and the Cure Autism Now Foundation (to A.R.).

1. Friedman C, Kra P, Yu H, Krauthhammer M, Rzhetsky A (2001) *Bioinformatics* S74–S82.
2. Rzhetsky A, Iossifov I, Koike T, Krauthhammer M, Kra P, Morris M, Yu H, Duboue PA, Weng W, Wilbur WJ, *et al.* (2004) *J Biomed Inform* 37:43–53.
3. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, *et al.* (2005) *Cell* 122:957–968.
4. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, *et al.* (2005) *Nature* 437:1173–1178.
5. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, *et al.* (2006) *Nat Genet* 38:285–293.
6. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, *et al.* (2003) *Science* 302:1727–1736.
7. Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) *Nature* 417:399–403.
8. Stibius KB, Sneppen K (2007) *Biophys J* 93:2562–2566.
9. Bader GD, Hogue CW (2002) *Nat Biotechnol* 20:991–997.
10. Jimenez-Sanchez G, Childs B, Valle D (2001) *Nature* 409:853–855.
11. Blake JA, Eppig JT, Bult CJ, Kadin JA, Richardson JE, Group MGD (2006) *Nucleic Acids Res* 34:D562–D567.
12. Dorogovtsev SN, Mendes JFF (2003) *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford Univ Press, Oxford).
13. Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) *Nature* 411:41–42.
14. Coulomb S, Bauer M, Bernard D, Marsolier-Kergoat MC (2005) *Proc Biol Sci* 272:1721–1725.
15. Ng PC, Henikoff S (2002) *Genome Res* 12:436–446.
16. George RA, Liu JY, Feng LL, Bryson-Richardson RJ, Fatkin D, Wouters MA (2006) *Nucleic Acids Res* 34:e130.
17. Smith NG, Eyre-Walker A (2003) *Gene* 318:169–175.
18. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, *et al.* (2004) *Proc Natl Acad Sci USA* 101:6062–6067.
19. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, *et al.* (2003) *Nature* 421:231–237.
20. Lopez-Bigas N, Blencowe BJ, Ouzounis CA (2006) *Bioinformatics* 22:269–277.
21. Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I, Marcotte EM (2004) *Curr Opin Struct Biol* 14:292–299.
22. Ramani AK, Bunescu RC, Mooney RJ, Marcotte EM (2005) *Genome Biol* 6:R40.
23. Arkinq DE, Chuqh SS, Chakravarti A, Spooner PM (2004) *Circ Res* 94:712–723.
24. Krauthhammer M, Kauffman CA, Gilliam TC, Rzhetsky A (2004) *Proc Natl Acad Sci USA* 101:15148–15153.
25. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi A-L (2007) *Proc Natl Acad Sci USA* 104:8685–8690.