

# Supporting Information on “The Human Disease Network”

Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal & Albert-László Barabási

## Contents

- S1. Construction of the diseasome map.
- S2. Properties of the HDN and DGN.
- S3. Component size distributions of HDN and DGN.
- S4. Genetic heterogeneity and connectivity of disorder classes.
- S5. Protein-protein interaction data.
- S6. Random control for the PPI-GDN overlap.
- S7. Gene Ontology analysis.
- S8. Gene expression microarray data.
- S9. Tissue homogeneity.
- S10. Random controls for gene expression analysis.
- S11. Mouse phenotype data.
- S12. Significance analyses of Fig. 4.
- S13. Centrality of somatic cancer genes.
- S14. Testing the robustness of the results: Analysis of the extended dataset.
- S15. Supporting references.

## Additional files:

- I. Table S1. Curated Morbid Map file (21, December 2005 version).
- II. Table S2. Network characteristics of human diseases.
- III. Table S3. Network characteristics of disease genes.
- IV. Table S4. List of human protein-protein interactions.
- V. Figure S9. Poster layout of the whole diseasome bipartite graph. The details of this poster are best viewed if printed with 300% magnification, or increased 300% in the PDF viewer.

## **S1. Construction of the diseasome map**

The most complete and best-curated list of known disorder-gene associations is maintained in the Morbid Map (MM) of the Online Mendelian Inheritance in Man (OMIM) (S1). Each entry of the MM is composed of four fields, the name of the disorder, the associated gene symbols, its corresponding OMIM id, and the chromosomal location. We downloaded the MM file on 21 December 2005. Out of 4,043 MM entries, we selected 2,929 entries with the “(3)” tag, for which there is strong evidence that at least one mutation in the particular gene is causative to the disorder. We then parsed these 2,929 disorder terms into 1,286 distinct disorders by merging disease subtypes of a single disease, based on their given disorder names. For example, the eleven complementation groups of Fanconi anemia are merged into the disease “Fanconi anemia” and the two fibromatosis entries are merged into the disease “fibromatosis” [see Supporting Information (SI) Fig. S1]. The merging was done first automatically by running a string-match script and then each entry was verified manually. Each disease was then assigned a unique disease ID.

Subsequently we classified each disorder into 20 primary disorder classes manually, following the classification scheme shown in Fig. 2. The classification is based on the physiological system affected by the disorder. For example, 113 disorders constitute a “cancer” class and 41 disorders such as atherosclerosis and stroke constitute a “cardiovascular” class. When a disorder affects multiple systems we tried to assign it to a primary class based on which system was most affected; if a primary class was not evident then the disorder was placed into the “multiple” class. While diseases categorization is often a cause of heated debate, occasional misclassification will not affect our results. Disorders having distinct multiple clinical features are assigned to the “multiple” class, with 155 disorders in this category. For 31 disorders there was insufficient information available for a clear class assignment, thus we annotated these into a “unclassified” class. Therefore every disorder was annotated into one of 22 disorder classes.

Finally, each gene symbol was mapped onto an Entrez ID, generating the list of disease gene-disorder class associations available as SI Table S1.

Tables S2 and S3 summarize the properties of the diseasome map. For each disorder Table S2 tabulates the disorder class, the degree  $k$  in the HDN, the class-degree  $\kappa$  (number of distinct disorder classes it connects to), the size  $s$  (number of associated genes) and the list of the associated disease genes. For each gene Table S3 tabulates the class of its associated disorders, and the number and the list of disorders to which it is associated.

518	Fabry disease (3)	GLA	301500	Xq22	Metabolic
519	Facioscapulohumeral muscular dystrophy-1A (3)	FSHMD1A, FSHD1A	158900	4q35	Muscular
520	Factor H and factor H-like 1 (3)	HF1, CFH, HUS	134370	1q32	Hematological
520	Factor V and factor VIII, combined deficiency of, 227300 (3)	MCFD2	607788	2p21-p16.3	Hematological
520	Factor VII deficiency (3)	F7	227500	13q34	Hematological
520	Factor X deficiency (3)	F10	227600	13q34	Hematological
520	Factor XI deficiency, autosomal dominant (3)	F11	264900	4q35	Hematological
520	Factor XI deficiency, autosomal recessive (3)	F11	264900	4q35	Hematological
520	Factor XII deficiency (3)	F12, HAF	234000	5q33-qter	Hematological
520	Factor XIIIa deficiency (3)	F13A1, F13A	134570	6p25-p24	Hematological
520	Factor XIIIb deficiency (3)	F13B	134580	1q31-q32.1	Hematological
522	Familial Mediterranean fever, 249100 (3)	MEFV, MEF, FMF	608107	16p13	Immunological
523	Fanconi anemia, complementation group A, 227650 (3)	FANCA, FACA, FA1, FA, FAA	607139	16q24.3	multiple
523	Fanconi anemia, complementation group B, 300514 (3)	FAAP95, FAAP90, FLJ34064, FANB	300515	Xp22.31	multiple
523	Fanconi anemia, complementation group C (3)	FANCC, FACC	227645	9q22.3	multiple
523	Fanconi anemia, complementation group D1, 605724 (3)	BRCA2, FANCD1	600185	13q12.3	multiple
523	Fanconi anemia, complementation group D2 (3)	FANCD2, FANCD, FACD, FAD	227646	3p25.3	multiple
523	Fanconi anemia, complementation group E (3)	FANCE, FACE	600901	6p22-p21	multiple
523	Fanconi anemia, complementation group F (3)	FANCF	603467	11p15	multiple
523	Fanconi anemia, complementation group G (3)	XRCC9, FANCG	602956	9p13	multiple
523	Fanconi anemia, complementation group J, 609054 (3)	BRIP1, BACH1, FANCI	605882	17q22	multiple
523	Fanconi anemia, complementation group L (3)	PHF9, FANCL	608111	2p16.1	multiple
523	Fanconi anemia, complementation group M (3)	FANCM, KIAA1596	609644	14q21.3	multiple
524	Fanconi-Bickel syndrome, 227810 (3)	SLC2A2, GLUT2	138160	3q26.1-q26.3	Metabolic
526	Farber lipogranulomatosis (3)	ASAH, AC	228000	8p22-p21.3	Metabolic
527	Fatty liver, acute, of pregnancy (3)	HADHA, MTPA	600890	2p23	Metabolic
528	Favism (3)	G6PD, G6PD1	305900	Xq28	Metabolic
530	Fechtner syndrome, 153640 (3)	MYH9, MHA, FTNS, DFNA17	160775	22q11.2	multiple
531	Feingold syndrome, 164280 (3)	MYCN, NMYC, ODED, MODED	164840	2p24.1	multiple
532	Fertile eunuch syndrome, 228300 (3)	GNRHR, LHRHR	138850	4q21.2	Endocrine
535	Fibrocalculus pancreatic diabetes, susceptibility to (3)	SPINK1, PSTI, PCTT, TATI	167790	5q32	Gastrointestinal
537	Fibromatosis, gingival, 135300 (3)	SOS1, GINGF, GF1, HGF	182530	2p22-p21	Connective tissue disorder
537	Fibromatosis, juvenile hyaline, 228600 (3)	ANTXR2, CMG2, JHF, ISH	608041	4q21	Connective tissue disorder

Figure S1. **A subset of the Morbid Map, providing the disorder-disease gene associations.** We parsed the MM list to merge eleven complementation groups of Fanconi anemia into a single disorder (id: 523) and two entries of fibromatosis into a single disorder (id: 537). The same procedure was reported for all entries, mapping 2,929 disorder entries in the MM into 1,286 distinct human disorder IDs. The full list of such disorders is provided in Table S1. The original MM list consists of the second to the fifth columns.

## S2. Properties of the HDN and DGN

The layouts of the giant components of both HDN and DGN in Fig. 2 were generated by a simple force-directed algorithm, followed by a local rearrangement for visual clarity, while leaving the network's overall layout unperturbed.

In the HDN, the number of genes associated with a disorder,  $s$ , has a broad distribution (Fig. S2a), indicating that most disorders relate to a few disease genes, while a handful of disorders relate to dozens of genes (SI Table S2). The HDN exhibits a skewed degree ( $k$ ) distribution (Fig. S2b), with most disorders linked to only a few other disorders, while a few disorders (Table S2) represent hubs that are connected to a large number of distinct disorders.

In the complementary DGN, while the number of genes involved in multiple diseases decreases rapidly, several disease genes are involved in as many as ten disorders, representing major hubs in the network (Fig. S2d).

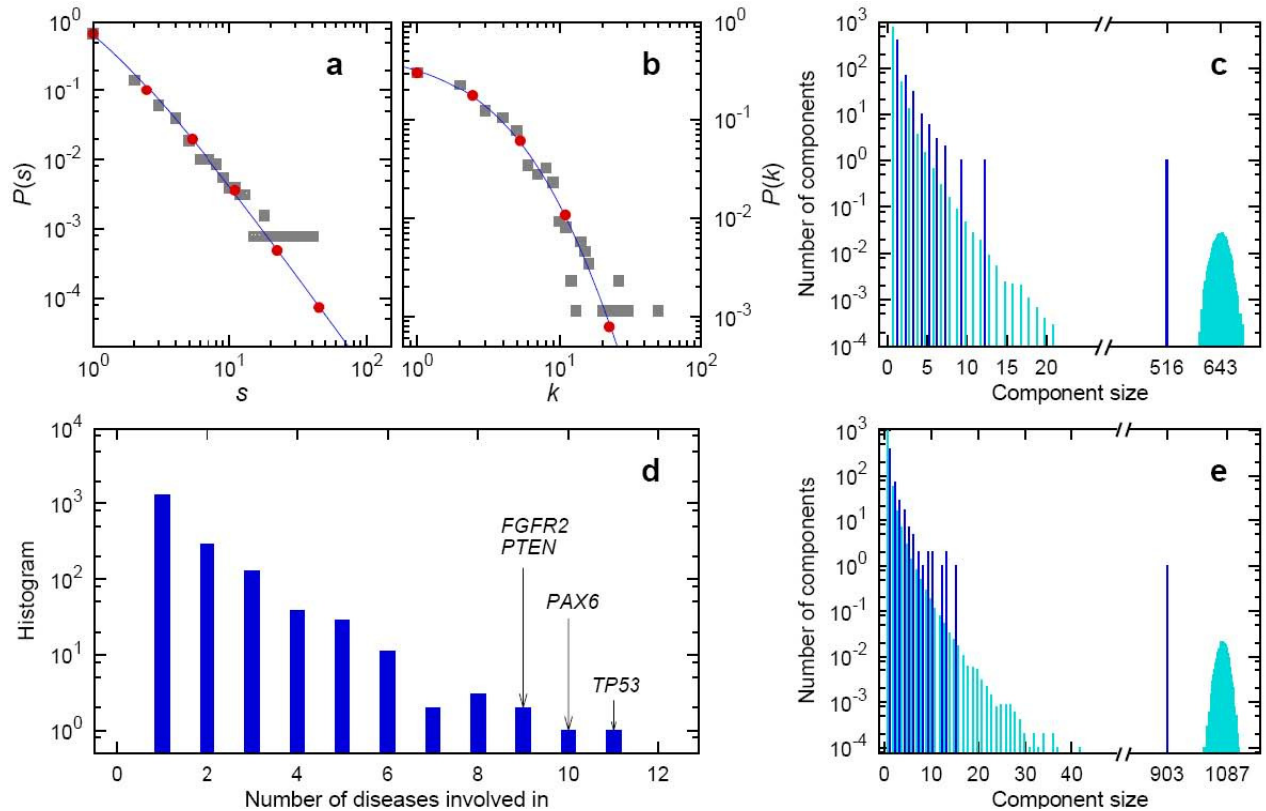


Figure S2. **Characterizing the topology of the HDN and the DGN.** (A-B) Distribution of (A) the size  $s$  (number of genes involved in a disorder) and (B) the degree  $k$  (number of other disorders a disorder shares genes with) of all disorders in the HDN. Gray symbols correspond to a linear binning, while red dots represent the logarithmically-binned data, maintaining the same statistical significance in each bin. The continuous lines represent the fit to the log-binned data, following the generalized power-law  $f(x)=c(x+a)^{-b}$  with (A)  $b \approx 2.7$  and (B)  $b \approx 6.5$ , obtained from the least-square fit. (C) Distribution of the cluster sizes in the HDN (blue), where the isolated peak at 516 corresponds to the size of the giant component. The component size distribution for the randomized network is also shown (light blue). (D) Histogram of the number of disorders a gene is involved in, identifying the four genes associated with the largest number of disorders. (E) Distribution of the connected component sizes in the DGN (blue). The largest component contains 903 genes. The component size distribution for randomized networks is also shown (light blue).

### S3. Component size distributions of HDN and DGN

The topology of the HDN and GDN networks deviates from random. To obtain random controls we randomly shuffled the disorder-disease gene associations, while keeping both the number of genes that a disorder is associated with and the number of disorders that a gene is implicated in unchanged. From these randomized disorder-disease gene associations we obtained the two

projections of randomized disease, the randomized HDN and the randomized GDN. To control for cluster size distribution, we generated  $10^4$  independent randomized samples.

The component size distributions, defined as the number of components with size  $s$ ,  $n(s)$ , for the two network projections obtained after the randomizations are shown in light blue in Fig. S2c and e. Both HDN and DGN have significantly smaller giant component than expected random, due to the functional clustering. Thus, HDN and DGN represent an intermediate structure between a completely randomized network with a very large giant component and a functionally fully segregated network which would be broken into isolated clusters, each representing a disorder class. Apart from the giant component, in both networks the sizes of disconnected components are distributed approximately as a power law, with the exponent about  $\approx -3$  in both cases. But a solid conclusion is difficult to be drawn due to the limited statistics.

#### **S4. Genetic heterogeneity and connectivity of disorder classes**

Genetic heterogeneity, specifically locus heterogeneity, means that mutations in more than one gene lead to similar disorder phenotypes (S2). We measured the genetic heterogeneity of a disorder class as the average number of genes in the disorders belonging to the selected class (*i.e.*, the average size of nodes in the class in the HDN). To quantify the statistical significance of the observed values, we randomized class annotations by randomly shuffling the disorder-class associations. According to the obtained  $P$ -values we identified significantly enriched (red) and significantly depleted (blue) classes (Fig. S3). The cancer and neurological disorder classes show high genetic heterogeneity, while the metabolic, skeletal, and multiple disorder classes show low genetic heterogeneity. We also calculated for each disorder class the fraction of disorders that are connected to each other in the HDN to quantify the “connectivity” of the particular disorder class. The statistical significance of the observed connectivity was assessed by the randomized disorder-disease gene associations (Fig. S3). By this measure, cancer is the most connected class and the metabolic disorder class is the least connected.

Percentage of disorders in giant component			
Percentage of disorders in the HDN			
Average number of genes			
CLASS			
Bone	1.8	80	63
Cancer	3.1	88	89
Cardiovascular	2.8	80	82
Connective tissue	2.1	71	75
Dermatological	2.2	77	57
Developmental	1.5	63	35
Ear,Nose,Throat	7.7	83	60
Endocrine	2.2	71	68
Gastrointestinal	1.7	61	29
Hematological	2.2	76	51
Immunological	1.9	57	62
Metabolic	1.5	43	34
Muscular	2.8	71	68
Neurological	2.6	67	71
Nutritional	6.0	75	100
Ophthalmological	2.7	77	81
Psychiatric	2.1	53	89
Renal	1.8	61	36
Respiratory	2.8	92	33
Skeletal	1.3	83	43
Multiple	1.6	71	49

Figure S3. **Genetic heterogeneity and connectivity of disorder classes.** Red (blue) denotes the enrichment (depletion) of the measure in the corresponding cell. Dim-colored cell denotes the observed value is not statistically significant ( $P > 10^{-2}$ ). The statistical significance of each value is calculated with the randomized HDN obtained from the randomized disorder-disease gene associations.

## S5. Protein-protein interaction data

To obtain a detailed human protein-protein interaction (PPI) data we combined two high quality systematic yeast two-hybrid experiments (S3, S4) with PPIs obtained from literature by manual curation (S3). The integrated set of PPIs contains 22,052 non-self-interacting, non-redundant interactions between 7,533 genes. The list of PPIs used is available as Table S4.

## S6. Random control for the PPI-GDN overlap

To generate the random control of the overlap between the PPI network and the GDN in Fig. 3a, we first identified 1,203 genes that are present both in the PPI network and the GDN and for each disorder  $i$  the number of genes  $n_i$  that are present in the PPI network. To obtain the random control of the overlap, we calculated for each disease  $i$  the number of PPIs between the  $n_i$  nodes selected randomly among the 1,203 genes while keeping the degree of the associated nodes. We performed this procedure for every disorder to obtain the number of all overlaps for a single random

configuration. We generated  $10^6$  independent random configurations to obtain significant statistics and  $P$ -value.

### S7. Gene Ontology analysis

If the HDN shows modular organization then a group of genes associated with the same common disorder should share similar cellular and functional characteristics, as annotated in Gene Ontology (GO; ref. S5). To investigate this, we measured the GO homogeneity (GH) of each disorder as the maximum fraction of genes in the same disorder that have the same GO terms. It is defined as

$$GH_i = \max_j [ n^j_i / n_i ],$$

where in this case  $n_i$  denotes the number of genes in the disorder  $i$  that have any GO annotations, and  $n^j_i$  the number of genes that have the specific GO term  $j$ . We calculated  $GH_i$  separately for each branch of GO, biological process (BP), molecular function (MF), and cellular component (CC). As expected, we find a significant elevation in the GH with respect to random controls in all three branches (Fig. S7). For example, we find a 23-fold increase in the perfect homogeneity for BP (79% vs. 3.4%), a 13-fold increase (75% vs. 5.5%) for MF, and a 9-fold increase (79% vs. 8.8%) for CC. To obtain the random control of the GO homogeneity distribution for each disorder we picked the same number of genes randomly in the GO annotation data and calculated their GO homogeneity. We generated  $10^4$  random instances to reach statistical significance.

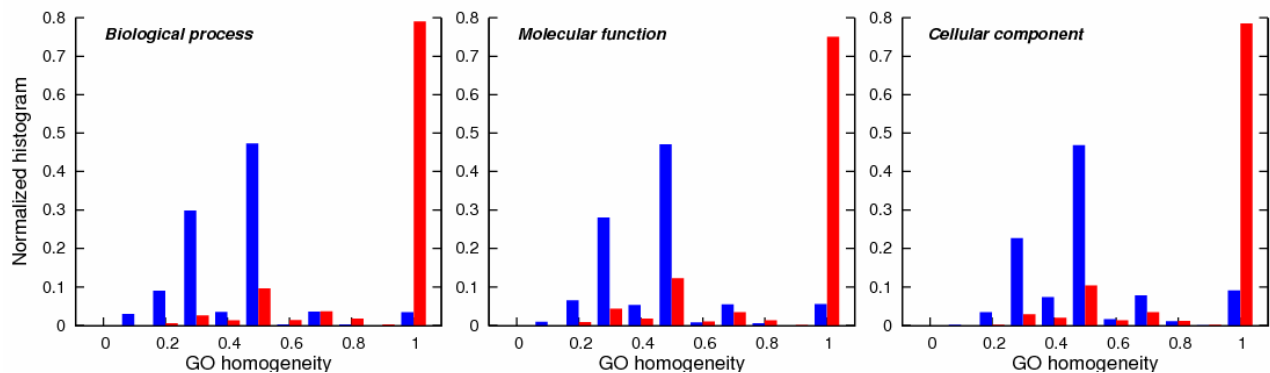


Figure S4. **Gene Ontology Homogeneity.** The GO homogeneity of disorders for the GO categories biological process (left), molecular function (middle), and cellular component (right). Red bars represent the actual histogram and the blue bars denote the random control, obtained for each disorder by choosing the same number of genes randomly.

### S8. Gene expression microarray data

To calculate the coexpression correlation between wild-type human gene transcripts, we used microarray data available for 36 normal human tissues (S6). By matching Entrez gene ID, 1,357 human disease genes had corresponding microarray probes (76% of human disease genes). A gene is considered to be “expressed” if the  $P$ -value associated with its transcript abundance is less than the threshold,  $P < 0.02$  (S6). We consider a gene as housekeeping gene if it is expressed in all 36 tissues. Genes that are not expressed in any examined tissues are excluded from the analysis.

### S9. Tissue homogeneity

The tissue homogeneity (TH) coefficient quantifies whether genes that are implicated in the same disorders tend to be expressed in similar human tissues. We define the TH of a disorder  $i$  as

$$TH_i = \max_j [n_i^j / n_i],$$

where  $n_i$  denotes the number of genes in the disorder  $i$  that are expressed in at least one tissue,  $n_i^j$  the number of genes that are expressed in the tissue  $j$  among them, and  $\max_j [\cdot]$  denotes the function returning the maximum-value argument across  $j$ . TH has the maximal value 1 if all the genes are expressed together in at least one tissue, and takes the minimum value  $1/n$  when all are expressed in different tissues. To obtain the random control of the tissue homogeneity distribution we picked the same number of genes randomly in the microarray data for each disorder and calculated their tissue homogeneity. We generated  $10^5$  random instances to reach statistical significance.

### S10. Random controls for gene expression analysis

To obtain the random control of the Pearson correlation coefficient (PCC) distributions for the gene expression in Fig. 3c and d, we calculated the distribution of all gene pairs in the microarray data (Fig. 3c) and the average PCC between the same number of genes chosen randomly from the microarray data (Fig. 3d). To obtain the  $P$ -values, we perform the  $\chi^2$ -test, calculating  $\chi^2$  values between the random normalized histograms obtained from the reference distributions (blue) and the actual distribution (red), performing  $10^6$  independent runs to obtain significant statistics.

### S11. Mouse phenotype data

To predict the essentiality of a human gene, we used the phenotype information of the corresponding mouse orthologs. A human gene was defined as “essential” if a knock-out of its mouse ortholog confers lethality. We obtained the human-mouse orthology and mouse phenotype

data from Mouse Genome Informatics (<http://www.informatics.jax.org>) on January 3, 2006. We considered the classes of embryonic/prenatal lethality and postnatal lethality as lethal phenotypes, and the rest of phenotypes as non-lethal ones. There were 1,267 mouse-lethal human orthologs, of which 398 have known human disease associations (22% of human disease genes).

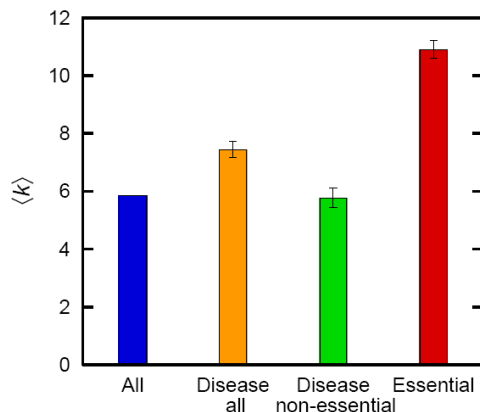


Figure S5. **Average degrees of groups of human genes.** Average degree of disease genes, essential genes, and non-essential disease genes, as well as all genes. Error bar denotes the standard error.

### S12. Significance analyses of the results in Fig. 4

To assess the statistical significance of the reported results, we apply the linear regression model and performed the  $\chi^2$  test for the significance of the measured trends dictated by the regression coefficient (S7). In particular, we take  $\log_2 k$  as the dependent variable in Fig. 4a, 4c, and 4d, which is more appropriate than  $k$  due to the power-law degree distribution. We found that the measured trends described by the linear regression model are statistically significant, with associated with P-values  $4.2 \times 10^{-6}$  (a),  $2.8 \times 10^{-5}$  (c),  $1.4 \times 10^{-4}$  (d),  $1.1 \times 10^{-16}$  (e), and  $3.5 \times 10^{-7}$  (f).

In addition, we assess the significance of the “strength” of the trends using the linear regression model. The slope  $A_O$  is obtained as the coefficient in the linear regression model and quantifies the strength (magnitude) of the trend. The  $P$ -value of the observed trend  $A_O$  will be the probability that we have  $A$  equal or larger (smaller, for the negative trend in f) than  $A_O$  purely by chance. To calculate the  $P$ -value, we randomized our sample by which we randomly redistribute the attributes (genes associated with diseases in a, for example) and perform the linear regression to obtain randomized (null) values of  $A$ , denoted by  $A_R$ . We found that  $A_R$  approximately follows a Gaussian distribution with zero mean and standard deviation  $\sigma_A$ . Thus, we can calculate the  $P$ -value

of the observed  $A_O$  from the Z-score defined as  $Z = (A_O - \langle A_R \rangle) / \sigma_A = A_O / \sigma_A$ . In Table S5, we summarize the results of the significance analysis according the described procedure.

Table S5. **Summary of the significance analysis for Fig. 4a, c-h** Fields in red denote significant positive trends, and those in blue do significant negative trends.

Panel	$A_O$	Z-score	P-value
4a	0.040	8.4	$1.6 \times 10^{-17}$
4c	0.092	8.6	$1.3 \times 10^{-17}$
4d	0.0088	2.1	0.015
4e	0.49	3.6	$1.7 \times 10^{-4}$
4f	-0.23	-5.4	$2.6 \times 10^{-8}$
4g	0.0064	8.1	$2.8 \times 10^{-16}$
4h	-0.0010	-4.7	$1.4 \times 10^{-6}$

### S13. Centrality of somatic cancer genes

The selection-based argument presented in the paper does not apply to the diseases that are caused by somatic mutations. Thus, for example, cancers caused by somatic mutations need not be at the functional periphery. Instead, given the severe physiological damage, often leading to death, resulted from such mutations, these mutations are expected to affect the functional center of the cell. To test this, we study the properties of genes whose somatic mutations are known to induce cancer. We obtain the list of somatic cancer genes from the Cancer Gene Census (<http://www.sanger.ac.uk/genetics/CGP/Census/>). From the analysis (Fig. S6) we find that these cancer genes indeed are (i) more likely to be encoded by hubs, (ii) show higher co-expression with the rest of the genes in the cell, and (iii) are more represented among housekeeping genes, confirming our expectation that somatic cancer genes are functionally central.

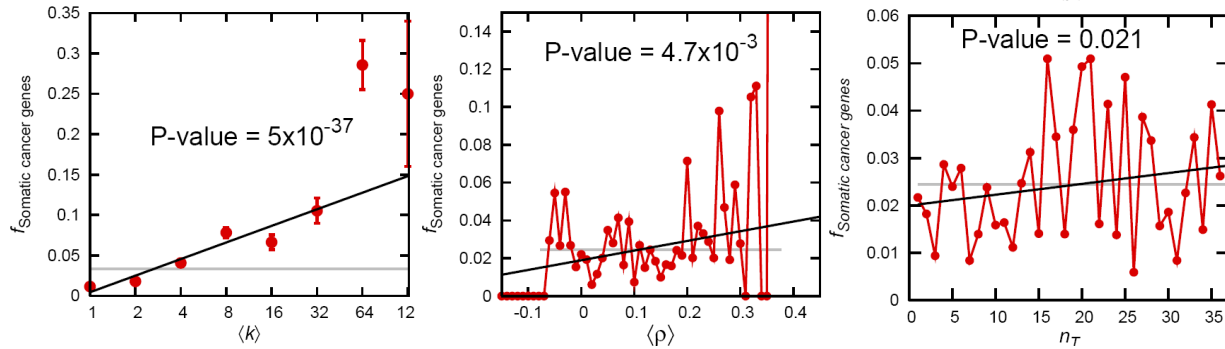


Figure S6. **Centrality of somatic cancer genes.** Plotted in each panel is the fraction of somatic cancer genes as a function of the average degree  $\langle k \rangle$  (left), the average co-expression coefficient  $\langle \rho \rangle$  (middle), and the number of tissues expressed  $n_T$  (right). All three quantities show positive trends, suggesting that somatic cancer genes are topologically and functionally central.

#### S14. Analysis of the extended dataset

In the analysis presented in the manuscript we included only the disorder-disease gene associations for which the wild-type gene is mapped and the mutation thereof is clearly demonstrated to be associated with the disorder, among all the data catalogued in the Morbid Map in OMIM (S1). In this section, we relax this criterion, and perform the same analysis with all the disorder-disease genes associations listed in the Morbid Map, including those with weaker evidences, for which only the mapping of either the wild-type gene [tag “(1)”] or the disease phenotype itself [tag “(2)”] is known (S1). Technically this extension did not require any further data curation, given that entries denoted by “(3)” tag in the OMIM database correspond to genes for which there is strong evidence that at least one mutation in the particular gene is causative of the disorder. In the earlier study we focused only on these genes (see Sec. I). In the extended study presented here, we included also those entries that have “(1)” or “(2)” tags. This extension will increase the coverage of the data but at the same time introduce potential errors in the form of disease genes that may not turn out to be associated with the specific disorder. The objective of this extension is to test the robustness of the main findings of current study to the increase in the coverage and the presence of noise in the dataset.

With this extension, we obtain a list of associations between 1,580 disorders and 2,765 disease genes, from 4,043 Morbid Map entries as of December 21, 2005. Thus the coverage in the genome increases by 50%. First we generate the analogue of Fig. 2a, the layout of the HDN (Fig. S7). While a slightly different clustering tool is used, giving a different overall appearance, the

overall layout of the disorder classes and the overall position of diseases remain largely unaltered. Next we perform the analysis analogous to that shown in Fig. 4. The overall behavior that i) the essential genes likely encode hubs whereas the non-essential disease genes are not particularly associated with hubs and ii) the highly coexpressed genes are depleted with the disease genes while enriched with essential genes is also manifest clearly (Fig. S8). Overall, the main findings of the current study are robust to the extension of the dataset which both increases the coverage and introduces potential errors.

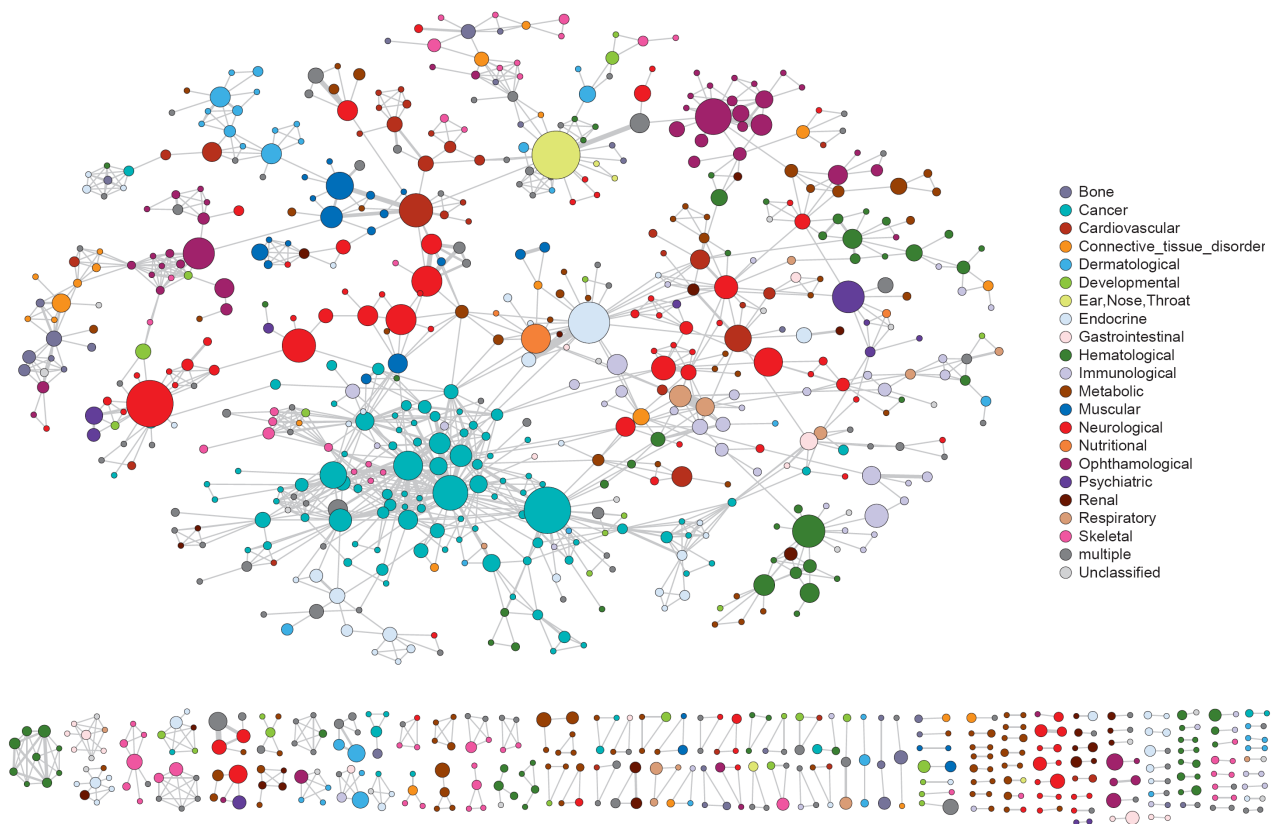
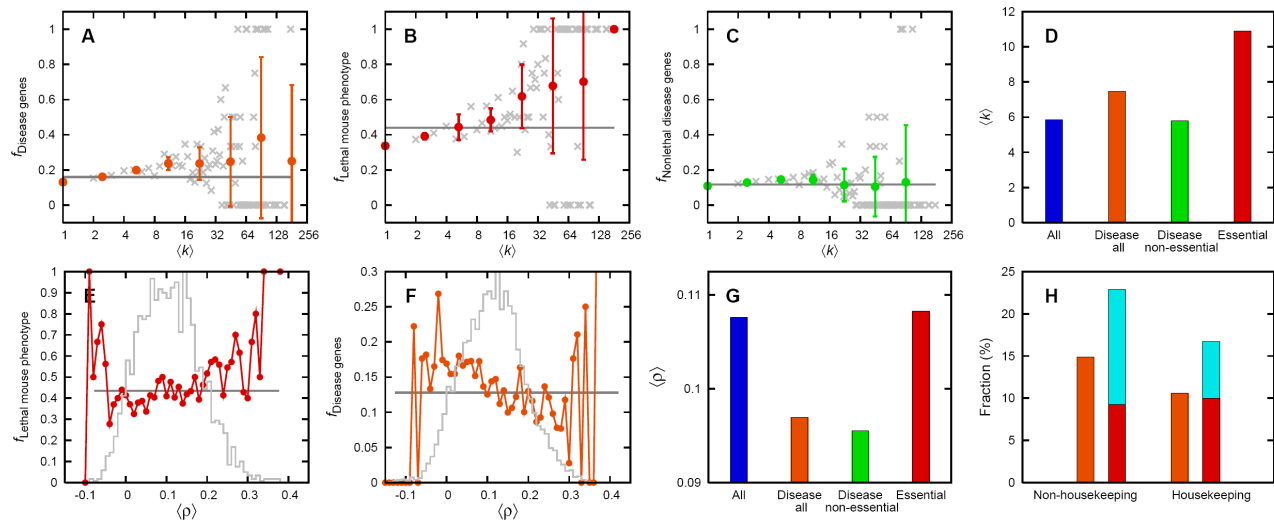


Figure S7. **Layout of the HDN with the extended dataset.** HDN with the extended dataset consists of 944 disorders with at least one link to other disorders and 576 disorders form the giant component.



**Figure S8. Functional characteristics of disease and essential genes for the extended dataset.** (a) The fraction of disease genes among those whose protein products interact with  $k$  other proteins. (b) The fraction of genes with lethal mouse phenotype among those whose protein products interact with  $k$  other proteins. (c) The same as in a, but excluding the proteins with lethal mouse phenotypes. (d) Average degree of disease, essential, and non-essential disease genes, as well as all genes (random). The fraction of essential genes (e) and disease genes (f) among those whose average PCC with other genes is  $\langle \rho \rangle$ . Gray horizontal lines in a-c and e-f indicate the global average and the gray bars in e-f show the number of genes (without scales) in each bin indicating that the high fluctuations at low and high  $\langle \rho \rangle$  are due to the small number of genes in those bins. (g) Average PCC of genes in different categories. (h) Fraction of disease genes (orange) and of genes with lethal (red) and non-lethal (light blue) mouse phenotypes among housekeeping and non-housekeeping genes.

## S15. Supporting references

- S1. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. (2005) *Nucleic Acids Res.* **33**, D514-D517.
- S2. Nussbaum, R. L., McInnes, R. R. & Willard, H. F. (2004) *Thompson & Thompson Genetics in Medicine*, 6<sup>th</sup> Ed. Saunders, Philadelphia, PA.
- S3. Rual, J.-F. *et al.* (2005) *Nature* **437**, 1173-1178.
- S4. Stelzl, U. *et al.* (2005) *Cell* **122**, 957-968
- S5. Gene Ontology Consortium (2006) *Nucleic Acids Res.* **34**, D322-D326.
- S6. Ge, X. *et al.* (2005) *Genomics* **86**, 127-141.
- S7. Daniel, W. W. (1991) *Biostatistics: A foundation for analysis in the health sciences*. Wiley.