Supplementary Materials for

Uncovering Disease-Disease Relationships Through The Human Interactome

Jörg Menche[1,2,3], Amitabh Sharma[1,2], Maksim Kitsak[1,2], Dina Ghiassian[1,2], Marc Vidal[2,4], Joseph Loscalzo[5], and Albert-László Barabási[1,2,3,5,*]

[1]Center for Complex Networks Research and Department of Physics, Northeastern University, 110 Forsyth Street, 111 Dana Research Center, Boston, MA 02115, USA.

[2]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, 450 Brookline Ave., Boston, MA 02215, USA

[3]Center for Network Science, Central European University, Nador u. 9, 1051 Budapest, Hungary

[4]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA.

[5]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 75 Francis Street, Boston, MA 02115, USA.

[*]To whom correspondence should be addressed; E-mail: alb@neu.edu.

This PDF file includes:

Supplementary Text

Figures S1 to S16

Table S1

Other Supplementary Materials for this manuscript:

Datasets S1 to S4 as a zipped archive

Python source code

Contents

List of Figures

List of Tables

Overview

In this document we present a detailed discussion of the datasets used and their analysis. In particular, we describe the different statistical tests and controls that we applied in the respective Sections. We also provide additional examples and extensions of the network-based methodology introduced in the main text. The document is organized as follows: In Section 1 we describe in detail all biological data we used and their analysis. In particular, we introduce the interactome sources in Section 1.1 and the gene-disease associations in Section 1.2. In Section 2 we discuss in more detail than in the main text the network-based measures of localization and separation. In Section 3 we compare the network-based overlap measure with gene-based measures. An analysis of the impact of biases and false positives on our main results is presented in Sections 4 and 5. In Section 6 we derive our main results from percolation theory. We briefly discuss the relationship between network communities and disease modules in Section 7 and introduce the layout algorithm we developed to visualize the disease-space in Section 8. Section 9 presents a discussion of several interesting overlapping disease pairs, including pairs without previously recognized shared disease genes. In Section 10 we describe how our network-based methodology can be used to enhance the interpretation of GWAS results. Section 11 summarizes the Supplementary data that we share with the community with this publication.

1 Data Compilation and Analysis

1.1 Interaction network data

1.1.1 The Human Interactome

In building the interactome, we rely only physical protein interactions with experimental support, hence we do not include interactions extracted from gene expression data or evolutionary considerations. In order to obtain an interactome as complete as currently feasible, we combine several databases with various kinds of physical interactions:

(i) Regulatory interactions: We use the TRANSFAC database [51] that lists interactions derived from the presence of a transcription factor binding site in the promoter region of a certain gene. The resulting network consists of 271 transcription factors regulating 564 genes via 1,335 interactions.

(ii) Binary interactions: We combine several yeast-two-hybrid high-throughput datasets [12, 14,

52–55] with binary interactions from IntAct [56] and MINT databases [57]. The sum of these data sources yields 28,653 interactions between 8,120 proteins. Note that IntAct and MINT provide interactions derived from both literature curation and direct submissions.

(*iii*) Literature curated interactions: These interactions, typically obtained by low throughput experiments, are manually curated from the literature. We use IntAct [56], MINT [57], BioGRID [58] and HPRD [59], resulting in 88,349 interactions between 11,798 proteins.

(*iv*) Metabolic enzyme-coupled interactions: Two enzymes are assumed to be coupled if they share adjacent reactions in the KEGG and BIGG databases. In total, we use 5,325 such metabolic links between 921 enzymes from [60].

(*v*) Protein complexes: Protein complexes are single molecular units that integrate multiple gene products. The CORUM database [61] is a collection of mammalian complexes derived from a variety of experimental tools, from co-immunoprecipitation to co-sedimentation and ion exchange chromatography. In total, CORUM yields 2,837 complexes with 2,069 proteins connected by 31,276 links.

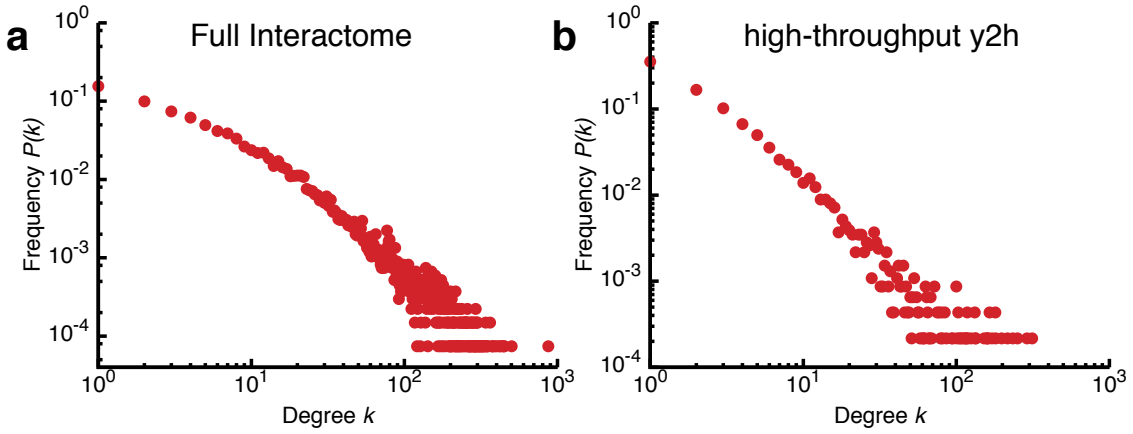
(*vi*) Kinase network (kinase-substrate pairs): Protein kinases are important regulators in different biological processes, such as signal transduction. PhosphositePlus [62] provides a network of peptides that can be bound by kinases, yielding in total 6,066 interactions between 1,843 proteins.

(*vii*) Signaling interactions: The dataset from [63] provides 32,706 interactions between 6,339 proteins that integrate several sources, both high-throughput and literature curation, into a directed network in which cellular signals are transmitted by proteins-protein interactions.

The union of all interactions obtained from (*i*)-(*vii*) yields a network of 13,460 proteins that are interconnected by 141,296 physical interactions. Note that we treat the interactome as an undirected network (see Section 2.3 for a discussion of the impact of directionality). The interactome is approximately scale-free (Figure S1a) and shows other typical characteristics as observed previously in many other biological networks [64, 65], such as high clustering or short pathlengths (Figure S1c)

1.1.2 Unbiased high-throughput Interactome

Since our interactome includes data from literature curation, it is inherently biased towards much studied disease-associated proteins and their interactions. We therefore complement our analysis



c

	full Interactome	high-throughput y2h
number of nodes N	13,460	4,612
number of links M	141,296	15,937
mean degree $\langle k \rangle$	21	7
mean clustering $\langle c \rangle$	0.17	0.05
mean shortest path $\langle l \rangle$	3.6	4.0
diameter $d_{\max}$	12	11

Figure S1: **Basic properties of the two interactomes used in this study. a,b,** Degree distribution $P(k)$ of the full interactome and the high-throughput yeast two-hybrid (y2h) network. **c,** Basic network properties.

using only interactions from well controlled and completely unbiased high-throughput yeast two-hybrid (y2h) datasets [12, 14, 53–55]. As detailed below (Sec. 6), these data are particularly suited to systematically address the effects of incompleteness, since all possible pairwise combinations of a given set of proteins have been tested in an unbiased fashion on the same platform. It contains a total of 4,612 proteins and 15,937 interactions and, taking its smaller size into account, shows expected characteristics such as degree distribution, clustering and pathlengths (Figure S1a,c).

1.2 Disease-gene associations

1.2.1 Data sources

We integrate two sources of disease-gene annotation:

OMIM: The OMIM database (Online Mendelian Inheritance in Man; `http://www.ncbi.nlm.nih.gov/omim`) [48] is a comprehensive collection covering all known diseases with a genetic component. The OMIM associations we use also include associations from UniProtKB/Swiss-Prot and have been compiled by [30].

GWAS: Results from GWAS (Genome-Wide Association Studies), provide unbiased, i.e., not hypothesis driven disease associations. However, the associations are often difficult to interpret, since many identified polymorphisms cannot be immediately linked to changes of a gene. The disease-gene associations from GWAS are obtained from the PheGenI database (Phenotype-Genotype Integrator; `http://www.ncbi.nlm.nih.gov/gap/PheGenI`) [31] that integrates various NCBI genomic databases. We use a genome-wide significance cutoff of p-value $\leq 5 \times 10^{-8}$.

1.2.2 Combining GWAS and OMIM

GWAS and OMIM yield lists of genes and associated diseases. However, gene and disease annotations and nomenclature are not standardized and furthermore the levels of disease specification are very heterogeneous. Some are very generic, such as "gene A is associated with *cancer*", while others refer to more specific disorders, e.g. *prostatic neoplasms*. Typically only the most specific associations are explicitly reported in the respective database. In order to combine OMIM and GWAS we use the MeSH (Medical Subject Headings; `http://www.nlm.nih.gov/mesh/`) vocabulary (Figure S2). Using the hierarchical structure of the MeSH classification, we can find all implicit associations by expanding from a given specific terms upward to the most general ones. For example, OMIM reports an association between the gene MYEOV and *prostatic neoplasms*. Using the disease category of the MeSH tree (i.e., the "C"-branch), we infer the implicit associations to the entire set of more general disease categories, such as *male urogenital diseases* and *neoplasms* (Figure S2a). In this way we can merge OMIM and GWAS, regardless of the level of disease specificity at which a particular association has been reported (Figure S2b). In total, we obtain 1,489 different diseases and 3,176 associated genes. Filtering out diseases with less than 20 associated genes, we are left with 299 diseases and 3,173 genes. Our interactome has connectivity information for 2,436 of the corresponding proteins, the remaining are disregarded.

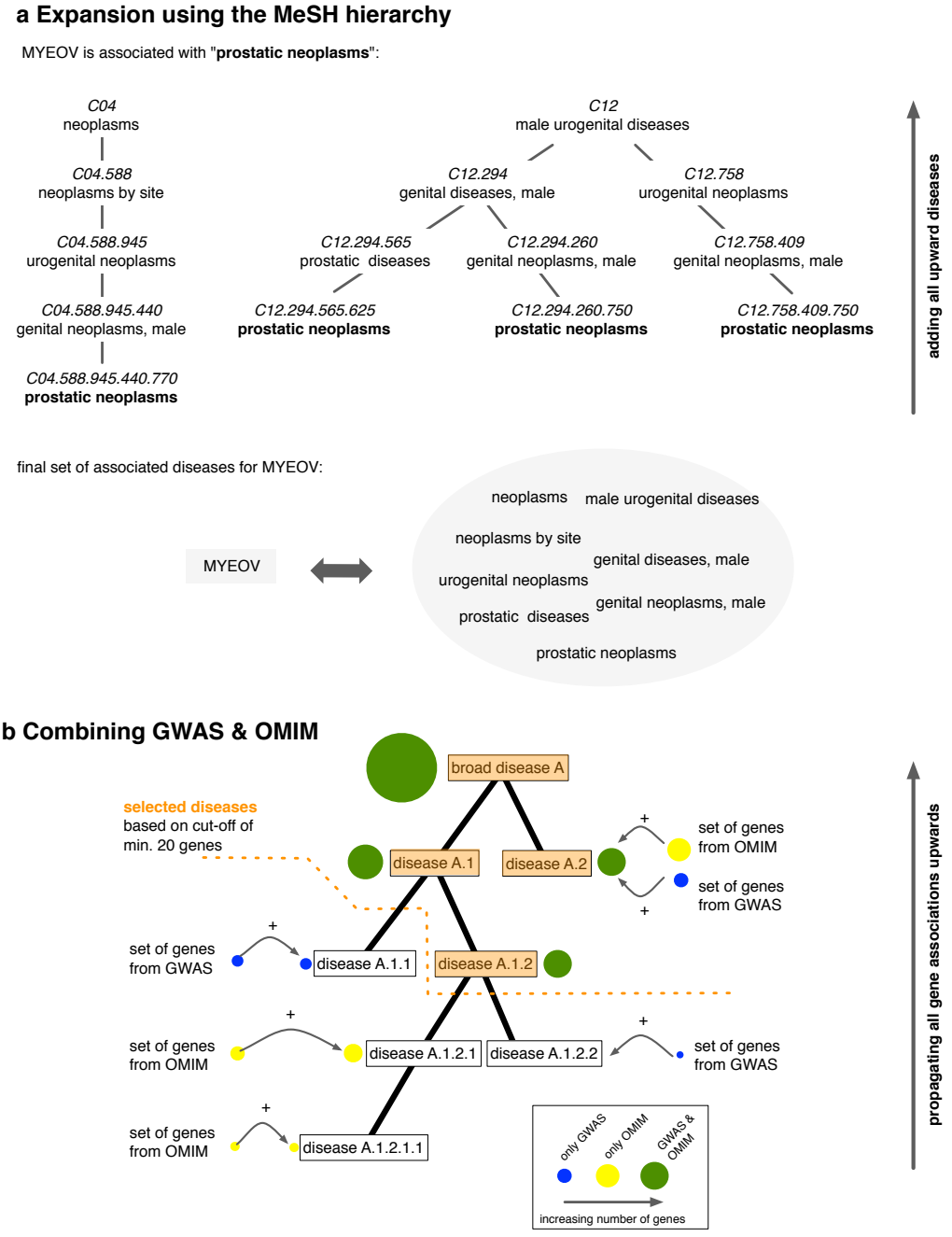


Figure S2: **Expansion of disease-gene associations using the MeSH hierarchy. a,** Example how the reported association of the gene MYEOV with prostatic neoplasms is expanded to all more general diseases, yielding a final set of eight associated disease. **b,** Illustration how gene-disease associations from OMIM and GWAS are combined along the MeSH hierarchy. Our final corpus consists of 299 diseases for which we find at least 20 associated genes.

1.3 Gene Ontology (GO)

1.3.1 Data source

GO annotations [49] for all genes are extracted from [`http://www.geneontology.org/`, downloaded Nov. 2011]. We only use high confidence annotations associated with the evidence codes `EXP, IDA, IMP, IGI, IEP, ISS, ISA, ISM` or `ISO`. In particular, we exclude annotations inferred from physical interactions (evidence code `IPI`) in order to avoid circularity in the evaluation of the GO-based similarity of proteins within the interactome. Following the curation process from [66], we further remove all the associations that have a non-empty "qualifier" column. The original GO files only contain the most specific annotations explicitly. In a last step, we therefore add all implicit more general annotations by up-propagating the given annotations along the full GO tree. Note that the strict filtering procedure reduces the total number of GO terms considerably, in the end we use only $\sim 50\%$ of all available terms for *biological processes* and *molecular function* and 25% for *cellular component*.

1.3.2 Similarity of disease genes

We quantify the functional similarity of genes by the *specificity* of their shared GO annotations, assuming that genes sharing very specific functions are more similar to each other than those who only share generic annotations. The specificity of a GO term i is measured by the total number of genes n_i annotated to it in the entire GO corpus. The similarity $\mathcal{S}_{\rm GO}(a,b)$ of two proteins a and b is then given by the most specific GO term they share:

$$\mathcal{S}_{\rm GO}(a,b) \equiv \frac{2}{\min(n_i)}\,. \tag{S1}$$

The value of $\mathcal{S}_{\rm GO}(a,b)$ ranges from $\mathcal{S}_{\rm GO}(a,b) \equiv 0$ for no shared GO terms, to $\mathcal{S}_{\rm GO}(a,b) = 1$ if a and b are the only two genes annotated to a specific GO term. The overal functional similarity of a set of genes associated with a particular disease is measured by the average $\mathcal{S}_{\rm GO}(a,b)$ over all $n_{\rm pairs}$ pairs of disease-associated proteins:

$$\langle \mathcal{S}_{\rm GO} \rangle = \frac{1}{n_{\rm pairs}} \sum_{\{a,b\}} \mathcal{S}_{\rm GO}(a,b). \tag{S2}$$

To test whether the functional annotations of disease proteins are more similar than expected for randomly chosen proteins, we compare the distribution $P(\mathcal{S}_{\rm GO}(a,b))$ measured for disease-associated protein pairs with the appropriate null distribution $P_{\rm rand}(\mathcal{S}_{\rm GO})$ of all protein pairs within the network. The statistical significance of an observed difference in the respective means $\langle \mathcal{S}_{\rm GO} \rangle$

and $\langle \mathcal{S}_{\rm GO}^{\rm rand} \rangle$ is given by the p-value from a Mann-Whitney U test. The whiskers in Figure 2c-h of the main text indicate the 5th, 25th, 50th, 75th and 95th percentiles of the data in the respective bins. In addition to the statistical significance we also determine the effect size using Glass's Δ

$$\Delta \equiv \frac{\langle \mathcal{S}_{\rm GO} \rangle - \langle \mathcal{S}_{\rm GO}^{\rm rand} \rangle}{\sigma \left(\mathcal{S}_{\rm GO}^{\rm rand} \right)} \,, \tag{S3}$$

where $\langle \mathcal{S}_{\rm GO}^{\rm rand} \rangle$ and $\sigma \left(\mathcal{S}_{\rm GO}^{\rm rand} \right)$ denote the mean and the standard deviation of the random distribution $P_{\rm rand}(\mathcal{S}_{\rm GO})$. Glass's Δ compares the observed and the random distribution, so that for example $\Delta = 1$ indicates that the observed mean is one standard deviation larger than the mean of random expectation. Since two distributions are being compared to each other, already relatively moderate Δ values indicate highly significant differences. Figure S3 gives the corresponding Δ values for the significances reported in Figure 2c-h of the main text. Again, we find that highly localized diseases (compare with Section 2 for the topological localization measure) exhibit strongly increased functional similarity.

1.3.3 Similarity of disease pairs

The overall functional similarity of the genes associated with two diseases A and B is determined as in Eq. (S2) by averaging over all pairs of proteins a and b with $a \in A$ and $b \in B$ under the condition $a \neq b$. This condition ensures that for diseases with common genes we do not take pairs into account where a gene would be compared to itself. We have also explicitly confirmed that the inclusion of such pairs would not lead to any noticeable differences for our purposes, as their contribution to the overall similarity is typically very limited. For example, for two diseases with 100 associated genes each, of which 10 are shared, Eq. (S2) involves a total of $100 \times 100 = 10,000$ gene pairs, of which only 10 are comparisons of a gene to itself.

The whiskers in Figure 3d-f of the main text indicate the 5th, 25th, 50th, 75th and 95th percentiles of the data in the respective bins. The global random expectation shown by the gray line gives the mean value of the full distribution of all pairwise gene similarities, corresponding to the null hypothesis of completely randomly assigned disease-gene associations.

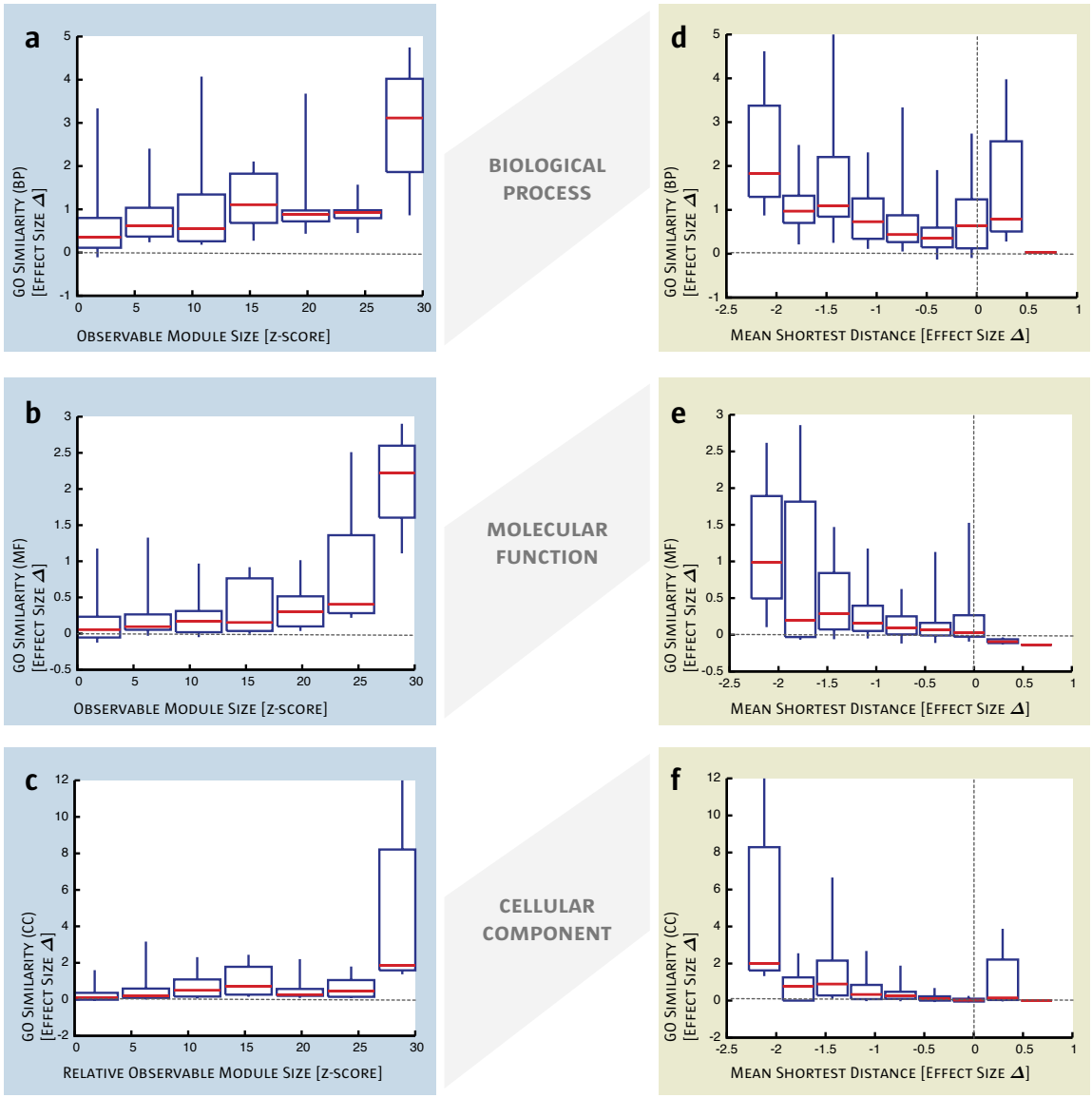


Figure S3: **Topological localization and biological** similarity of disease genes. The degree of the network-based localization of a disease compared to randomly distributed proteins is measured by the z-score of its observable module size S_i and the effect size Δ of the mean shortest distance $\langle d_s \rangle$. Random expectation is given by z-score $= \Delta = 0$ (indicated by the dotted lines). The biological similarity of the disease proteins is determined from their gene ontology annotations: For each disease, we compute how similar its associated genes are in terms of their biological processes (a,d), molecular function (b,e) and cellular component (c,f) and compare the result to random expectation using the effect size Δ. Most diseases show functionally more similar genes than expected ($\Delta > 0$), yet the effect is much more pronounced for diseases that are topologically localized. We find that the more localized a disease is topologically, i.e., the larger the z-score of S_i or the smaller Δ for $\langle d_s \rangle$, the more similar are the associated genes.

1.4 Gene co-expression analysis

1.4.1 Data source

We use tissue specific gene expression data from [36]. Of the 79 tissues in the dataset, we exclude five cancerous tissues[1] and four fetal tissues[2], leaving only data from healthy tissues.

1.4.2 Similarity of disease genes

In order to quantify the extent to which genes associated with the same disease are co-expressed, we calculate the Spearman correlation coefficient $\rho(a, b)$ and the corresponding p-value for each pair of genes a and b across all 70 tissues and average over all n_{pair} pairs:

$$\langle \rho \rangle \equiv \frac{1}{n_{\text{pair}}} \sum_{\{a,b\}} |\rho(a, b)| \,. \tag{S4}$$

For genes with multiple transcripts we use the transcript with the highest expression level. The Spearman correlation coefficient is often preferred over other measures like Pearson's r or Kendall's τ as it is more robust to outliers. For our purposes, however, we find that all three measures give comparable results.

1.4.3 Similarity of disease pairs

The expression similarity of the genes associated with two diseases A and B is computed by averaging $|\rho(a, b)|$ over all pairs of proteins a and b with $a \in A$ and $b \in B$. As for the GO similarity of disease pairs (Section 1.3.3) we only use $a \neq b$ gene pairs to ensure that for diseases with common genes we do not compare a gene to itself. The whiskers in Figure 3g of the main text indicate the 5th, 25th, 50th, 75th and 95th percentiles of the data in the respective bins. The global random expectation shown by the gray line gives the mean co-expression of all protein pairs in the network.

1.5 Symptom similarity

We use data from [38], where symptom-disease associations were extracted from large-scale medical bibliographic records in PubMed and the related MeSH metadata. Therein, each disease j is represented by a vector $\vec{d_j} \equiv (w_{1,j}, w_{2,j}, \ldots, w_{n,j})$, in which $w_{i,j}$ quantifies the PubMed co-occurrence

[1] `Leukemia.chronicMyelogenousK.562`, `Leukemia.promyelocytic.HL.60`, `Leukemialymphoblastic.MOLT.4` `Lymphoma.burkitt.s.Daudi` & `Lymphoma.burkitt.s.Raji`

[2] `FetalThyroid`, `Fetalbrain`, `Fetalliver`, `Fetallung`

of disease j with symptom i. The symptom similarity of two diseases A and B is measured by the Cosine similarity c_{AB} of their respective symptom vectors $\vec{d_A}$ and $\vec{d_B}$

$$c_{AB} \equiv \frac{\sum_{i=1}^{n} w_{i,A}\, w_{i,B}}{\sqrt{\sum_{i=1}^{n} w_{i,A}^2}\sqrt{\sum_{i=1}^{n} w_{i,B}^2}}\,, \tag{S5}$$

such that $c_{AB} = 0$ if A and B have no common symptoms and $c_{AB} = 1$ for diseases with identical symptoms. The whiskers in Figure 3h of the main text indicate the 5th, 25th, 50th, 75th and 95th percentiles of the data in the respective bins. Random expectation is given by the mean of the distribution of all disease pairs.

1.6 Comorbidity

We use a large Medicare patient medical history dataset [39, 60] to analyse the comorbidity of disease pairs, i.e., the tendency of two diseases to co-occur in the same patients. The data contains 13,039,018 patients diagnosed with one or more diseases over a period of 4 years. The patients are over 65 years old, mainly white and with a fraction of 58.3% of women. Comorbidity is quantified by the relative risk RR:

$$\begin{aligned} RR &= \frac{n_{AB} n_{\text{tot}}}{n_A n_B}\,, \text{ where} \\ n_{\text{tot}} &= \text{total number of patients in the data} \\ n_{A,} n_B &= \text{number of patients diagnosed with disease } A \text{ and } B\text{, respectively} \\ n_{AB} &= \text{number of patients diagnosed with both diseases } A \text{ and } B. \end{aligned} \tag{S6}$$

The whiskers in Figure 3i of the main text indicate the 5th, 25th, 50th, 75th and 95th percentiles of the data in the respective bins. A relative risk $RR > 1$ between two diseases indicates that they are diagnosed more often in the same patients than expected by chance given their individual prevalences. To evaluate the statistical significance of an obtained value of RR, we determine the lower and upper bounds (b_l and b_u) of the 99% confidence interval as in [67]:

$$\begin{aligned} b_{l,u} &= RR \times \exp(\pm 2.45\,\sigma)\,, \text{ with} \\ \sigma &= \frac{1}{n_{AB}} + \frac{1}{n_A n_B} - \frac{1}{n_{\text{tot}}} - \frac{1}{n_{\text{tot}}^2}\,. \end{aligned} \tag{S7}$$

The diagnoses in the database are given as ICD9 codes (International Statistical Classification of Diseases and Related Health Problems), which we manually mapped to the corresponding MeSH term. Similarly to the MeSH hierarchy discussed above in Section 1.2, we include for any given ICD9 code also all patients diagnosed with any more specific IC9 code.

1.7 Biological pathways

We use the Molecular Signatures Database (MSigDB) published by the Broad Institute, Version 3.1 [50]. MSigDB integrates several pathway databases, we use the ones from KEGG, Biocarta and Reactome. The enrichment analysis between a given gene set and a pathway is done using Fisher's exact test. The reported p-values are adjusted for the number of tested pathways (Bonferroni correction).

2 Network Localization

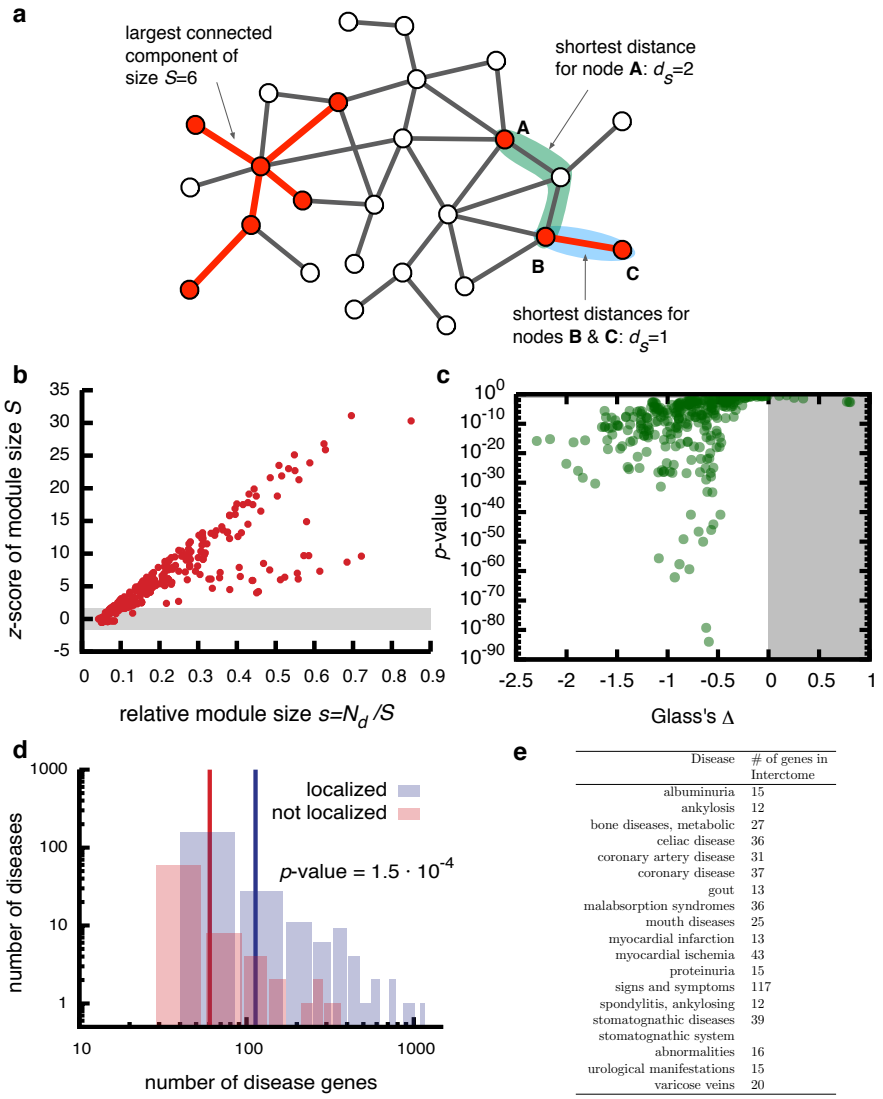
2.1 Measures of localization

We use two complementary measures to quantify the degree to which disease proteins tend to agglomerate in localized interactome neighborhoods (Figure S4a):

Module Size S. The first measure is given by the size of the largest connected module S, i.e., the highest number of disease proteins that are directly connected to one another. In addition to its intuitive interpretation, we can apply tools from statistical physics to understand many of its properties analytically. It is, however, relatively sensitive to data incompleteness. In extreme cases, a single missing link in the interactome or a single protein, whose disease association is not known, may destroy the connected component and leave the proteins isolated. To further substantiate our hypothesis of the existence of disease neighborhoods we therefore complement our analysis by measuring the distribution of shortest distances d_s between disease genes.

Shortest Distance d_s: For each of the N_d disease proteins we determine the *shortest* distance d_s to the next closest protein associated with the same disease, resulting in a distribution $P(d_s)$ of N_d data points. The average value $\langle d_s \rangle$ can be interpreted as the diameter of a disease on the interactome.

There are several possible variations and extensions of this distance measure. In particular, we have explored using *all* pairwise distances instead of only the distance to the next closest protein. We find that while the general results do not depend on the exact choice, d_s is the most predictive quantity, offering higher effect size with statistical significance. The reason for this is that d_s more appropriately handles cases where a disease module is split into several "islands", for example due to network incompleteness. Whereas d_s correctly reflects the high degree of localization within the



e

Disease	# of genes in Interctome
albuminuria	15
ankylosis	12
bone diseases, metabolic	27
celiac disease	36
coronary artery disease	31
coronary disease	37
gout	13
malabsorption syndromes	36
mouth diseases	25
myocardial infarction	13
myocardial ischemia	43
proteinuria	15
signs and symptoms	117
spondylitis, ankylosing	12
stomatognathic diseases	39
stomatognathic system abnormalities	16
urological manifestations	15
varicose veins	20

Figure S4: **Network-based localization. a,** Illustration of the network-based localization measures. The disease proteins (red) form one connected cluster of six proteins, one cluster of two proteins and one protein without connection to other disease proteins. The size of the observable module is therefore given by $S = 6$. For three proteins, A,B and C, the respective shortest distance to the next closest other disease protein is shown. **b,** Statistical significance of the observed module as measured by the $z-$score and relative module size $s = S/N_d$ for all diseases. Diseases outside the gray area show statistically significant modules (in total 244 out of 299). **c,** Statistical significance (p-value) and effect size (Glass's Δ) of the localization as measured by the mean shortest distance for all disease. 263 out of 299 exhibit significantly shorter distances compared to random expectation. **d,** Distribution of the number of diseases and their respective number of disease proteins for localized and not localized diseases. **e,** Table with the 18 diseases that do not show significant localization according to either module size or shortest distance distribution.

individual islands, it is diluted when the distances of all pairs are averaged.

Statistical Analysis: To test whether disease proteins have a tendency to agglomerate in specific interactome neighborhoods, we need to compare the measured values of S and $\langle d_s \rangle$ with a suitable random control. The appropriate null model here is to randomize the disease associations of the proteins and distribute them uniformly on the network. We then determine the resulting largest component and shortest distances of these randomized protein sets. Repeating the procedure 100,000 times yields distributions $P^{\text{rand}}(S)$ and $P^{\text{rand}}(d_s)$, from which we compute the statistical significance of the real data. For the size of the connected component we use the z-score

$$z\text{-score} \equiv \frac{S - \langle S^{\text{rand}} \rangle}{\sigma\left(S^{\text{rand}}\right)}\,, \tag{S8}$$

where $\langle S^{\text{rand}} \rangle$ and $\sigma\left(S^{\text{rand}}\right)$ denote the mean value and standard deviation of the random expectation $P^{\text{rand}}(S)$. Assuming normality of $P^{\text{rand}}(S)$, which is a good approximation for giant component sizes [68], we can analytically calculate a corresponding p-value for each z-score, yielding a threshold of z-score $\gtrsim 1.6$ for modules to be larger than expected by chance with significance p-value ≤ 0.05. For the shortest distances we compare $P(d_s)$ to the respective random distribution $P^{\text{rand}}(d_s)$ using a Mann-Whitney-U test. We find that 244 of the 299 diseases have a statistically significant module size, 263 have significantly shorter distances, and 226 fulfill both criteria (Figure S4b,c). Figure S4d compares the number of disease genes of these 226 well localized diseases to the number of diseases genes of the remaining not localized diseases. The average size of the well-localized diseases is $S \approx 112$, twice the size of the non-localized diseases $S \approx 62$. This is in agreement with the predictions from percolation theory that smaller modules require a higher coverage in order to be observable. Figure S4e lists the diseases which do not show localization in either S_i or d_s. In addition to the purely statistical argument based on network coverage, we observe that the not localized diseases fall roughly into two categories: (*i*) Diseases that have been extensively studied, like *coronary artery disease*, *coronary disease*, *myocardial infarction*, and *myocardial ischemia*. All of these disorders may also encompass a rather large $\langle d_s \rangle$ given the many different disease mechanisms at play in them. (*ii*) Relatively poorly studied diseases, like *stomatognathic diseases* or *varicose veins*. An exception to the observation that not localized diseases have only few associated genes is *signs and symptoms*. While formally being categorized as a disease within the MeSH classification (MeSH tree number `C23.888`), it is an umbrella term, rather than a concrete disease, so we would not expect it to be well localized within the interactome.

Complementing the analysis of randomly distributed disease proteins, we also explored, whether

the network localization of disease proteins could be the result of a particular wiring structure intrinsic to scale-free networks such as the interactome: Using the configuration model [69, 70] we constructed 100,000 randomized versions of the interactome with fixed disease associations and constant degrees of the proteins, but randomized interaction partners (degree preserving randomization [71]). Again, we find the modules of 232 out of 299 diseases to be statistically significant, hence their localization cannot be attributed to structural network properties alone.

2.2 Network separation

We quantify the network-based separation s_{AB} of two diseases A and B by comparing the mean shortest distances $\langle d_{\mathrm{AA}}\rangle$ and $\langle d_{\mathrm{BB}}\rangle$ *within* the respective diseases, to the mean shortest distance $\langle d_{\mathrm{AB}}\rangle$ *between* their proteins:

$$s_{\mathrm{AB}} \equiv \langle d_{\mathrm{AB}}\rangle - \frac{\langle d_{\mathrm{AA}}\rangle + \langle d_{\mathrm{BB}}\rangle}{2}\,. \tag{S9}$$

Note that proteins associated with both diseases A and B contribute with $d_{\mathrm{AB}} = 0$, see Figure S5 for a detailed example. In general, s_{AB} is bound by $-d_{\max} \leq s_{\mathrm{AB}} \leq d_{\max}$, where $d_{\max} = 13$ denotes the diameter of the network, i.e., the maximal distance of all node pairs. For disease pairs with no common proteins the minimal value increases to $-d_{\max}+1$. If we consider only diseases with at least two associated proteins, the maximal value is given by $d_{\max}-1$. Note that the separation parameter s_{AB} is not at intensive quantity, i.e., its magnitude depends on the size (number of proteins) of the individual diseases. Very large values are therefore obtained for two small, well separated diseases. Given the finite network space, diseases with a high number of associated genes, which are the most explored diseases, cannot achieve too high s_{AB}. The current network incompleteness leads to further scattering for larger diseases and therefore reinforces this effect. Very small s_{AB} values, on the other hand, are obtained for large diseases with big gene overlap. This is mainly the case for disease pairs where one disease is a variant of or precursor to the other, such as *atherosclerosis* and *coronary artery disease* ($s_{\mathrm{AB}} = -2.1$), or pathobiologically related diseases with the same broad classification, such as *biliary tract diseases* and *hepatic cirrhosis* ($s_{\mathrm{AB}} = -1.3$). We take this as a proof of its self-consistency. The most interesting cases, with so far unknown relationships are found mostly in the range of relatively small negative s_{AB}. In Section 9 we discuss a few examples.

2.2.1 Randomization of disease-gene associations

We have seen above that the value of s_{AB} depends to some extent on the number of proteins associated with the diseases A and B. We therefore assess the statistical significance of each disease pair individually, using two complementary random models:

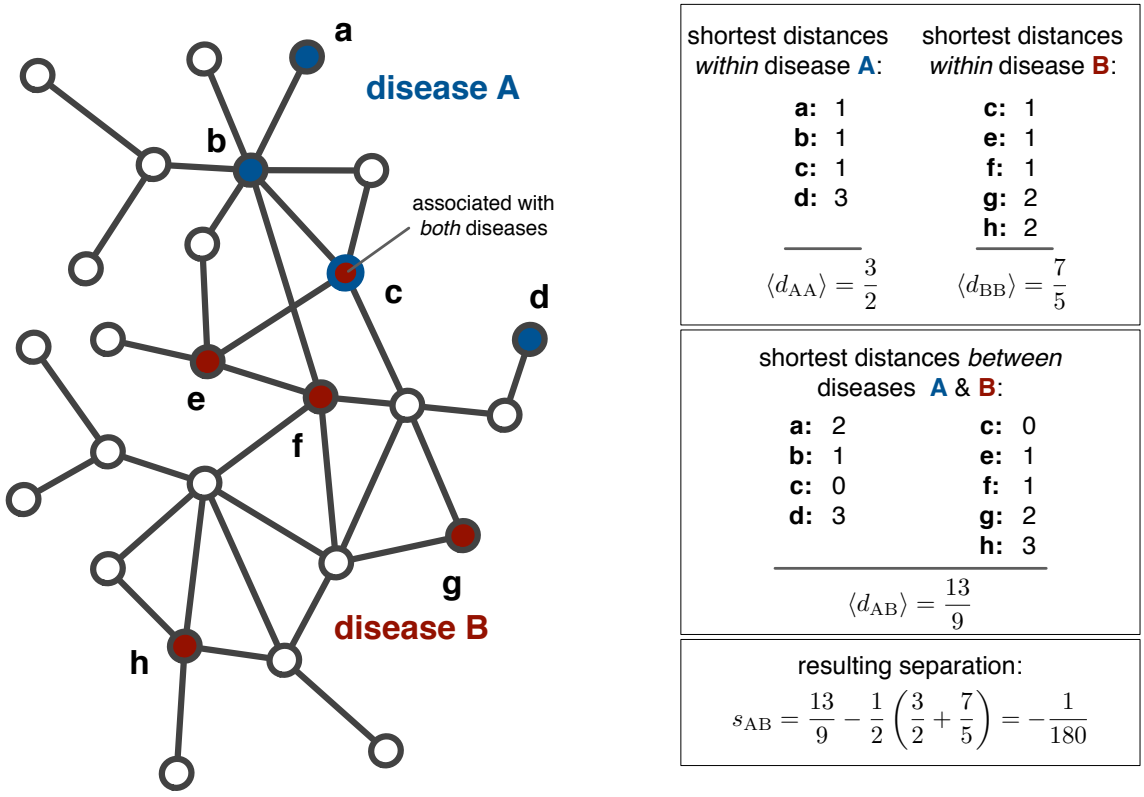


Figure S5: **Network-based separation.** Illustration of the network-based separation measure for two diseases A (blue) and B (red) with one shared protein ("c"). The tables on the right give the values of the mean shortest distances *within* the diseases, $\langle d_{\mathrm{AA}} \rangle$ and $\langle d_{\mathrm{BB}} \rangle$, as well as the distances for all protein pairs *between* them, $\langle d_{\mathrm{AB}} \rangle$.

Full randomization model. For two diseases A and B we draw the same number of proteins as in the respective sets of associated proteins completely at random and compute their corresponding separation $s_{\mathrm{AB}}^{\mathrm{rand}}$ on the network.

MeSH preserving randomization model. As detailed in section 1.2, our method to construct disease-gene association allows us to identify disease pairs with implicit correlations in their respective gene sets *via* the MeSH hierarchy. For example, as explained above, *prostate cancer* is by definition a complete subset of *neoplasms*. Even though the vast majority (98%) of the disease pairs do not involve such a relation, it is instructive to consider them separately using a randomization model that fully preserves all MeSH relationships between the diseases: Here, we only randomize the original disease-associations from OMIM and GWAS at the lowest level of the MeSH hierarchy and then perform the expansion along the MeSH tree in a second step. This procedure ensures

that the correlation structure imposed by the MeSH relation between certain diseases is kept intact.

For each disease pair, we performed 1,000 randomizations according to these two models and obtained $P(s_{\mathrm{AB}}^{\mathrm{rand}})$ as shown in Figure S6a,b for the same two disease pairs as in Figure 3a-c in the main text. Figure S6c,d shows for all disease pairs the observed s_{AB} *vs.* their corresponding mean random expectation $\langle s_{\mathrm{AB}}^{\mathrm{rand}} \rangle$ according to the two randomization schemes. We see that the full randomization yields values sharply centered around $s_{\mathrm{AB}}^{\mathrm{rand}} \approx 0$, whereas the real s_{AB} exhibit a much broader range of values. The MeSH preserving model, in contrast, also exhibits a number of randomized pairs with relatively large negative s_{AB} values. These are disease pairs that by definition share a considerable number of genes. Overall, however, their number is still very low as can be seen more clearly in the respective histogram (top panel of Figure S6d; note the logarithmic scale on the y-axis).

In order to quantify the difference between the observed value s_{AB} and random expectation $P(s_{\mathrm{AB}}^{\mathrm{rand}})$, we use the z-score

$$z\text{-score} \equiv \frac{s_{\mathrm{AB}} - \langle s_{\mathrm{AB}}^{\mathrm{rand}} \rangle}{\sigma\left(s_{\mathrm{AB}}^{\mathrm{rand}}\right)}\,, \tag{S10}$$

where $\langle s_{\mathrm{AB}}^{\mathrm{rand}} \rangle$ and $\sigma\left(s_{\mathrm{AB}}^{\mathrm{rand}}\right)$ denote the mean value and standard deviation of $P(s_{\mathrm{AB}}^{\mathrm{rand}})$. Negative (positive) z-scores imply that s_{AB} is smaller (larger) than expected by chance. Note that the two randomization models have some subtle differences regarding their interpretation, in particular for overlapping disease pairs: In the full randomization model a z-score < 0 indicates that the two respective diseases are more overlapping than expected by chance. Even though the underlying null hypothesis assumes completely independent gene sets, this conclusion also applies for disease pairs that are related *via* MeSH. Their shared genes contribute to the network overlap and drive s_{AB} towards negative values, so large negative s_{AB} correctly reflect their close relation. However, since the MeSH preserving randomization model retains the amount of shared genes, a positive z-score does not necessarily mean that an observed $s_{\mathrm{AB}} < 0$ is spurious. It rather indicates that the amount of shared genes dominates the network separation. In comparison to the fully randomized model, the MeSH preserving model therefore allows us to quantify how surprising an observed overlap is *beyond* a known MeSH relationship.

Significance analysis. Assuming normality of $P(s_{\mathrm{AB}}^{\mathrm{rand}})$, we can analytically calculate a corresponding p-value for each z-score, yielding a threshold of $|z\text{-score}| \gtrsim 1.6$ for a disease pair to be more/less overlapping than expected by chance with significance p-value ≤ 0.05. We also determined an

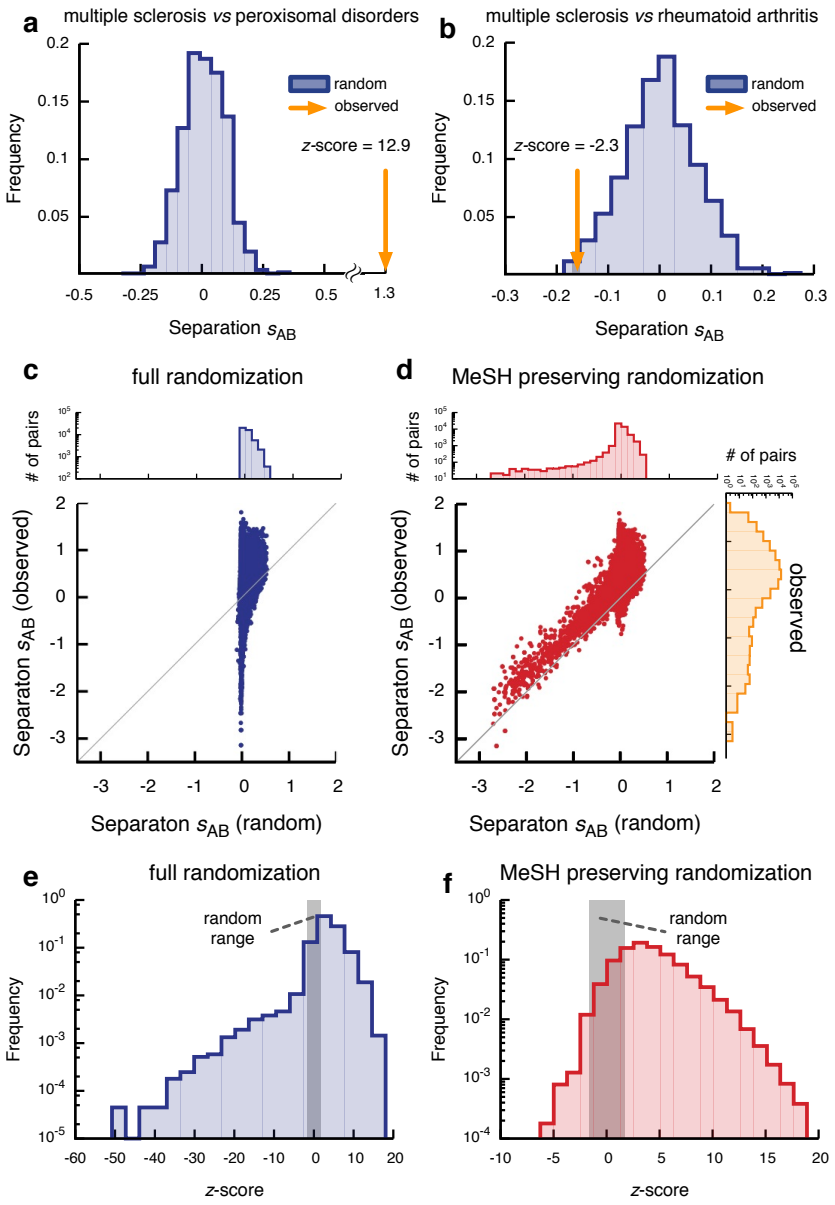


Figure S6: **Randomization of network separation.** **a,b,** Distribution of s_{AB}^{rand} values obtained from 1,000 random simulations for the two disease pairs shown in Figure 3a-c of the main text. The observed s_{AB} and the corresponding z-scores are indicated by the yellow arrow. **c,** Observed separation s_{AB} *vs.* random expectation according to the full randomization model for all disease pairs. The values for the random expectation represent mean values across 1,000 simulations. The histogram at the top gives the distribution of random s_{AB} values across all disease pairs. **c,** Same as **b**, but for the random expectation according to the MeSH-preserving randomization model. The histogram on the right gives the distribution of observed s_{AB} values across all pairs. **e,f,** Distribution of all z-scores of all disease pairs for the two randomization model. Gray areas indicate the range of random expectation $|z\text{-score}| \leq 1.65$ that corresponds to p-value ≥ 0.05.

empirical p-value for each pair by computing the fraction of random simulations in which $s_{\mathrm{AB}}^{\mathrm{rand}}$ exceeds the observed s_{AB} in the appropriate direction. We find that within the limitations imposed by the finite number of simulations the empirical p-values are in good agreement with the analytical estimate based on the z-score. Either measure can now be used to asses the statistical significance of an observed network separation. The two disease pairs in Figure S5a,b, for example, have z-score $= -2.3$ and z-score $= 12.9$, indicating that the diseases are significantly overlapping and separated, respectively (the diseases are unrelated within MeSH, so both randomization models yield similar results). Figure S6e,f shows the distribution of z-scores for all disease pairs. For both randomization models, the vast majority of disease pairs reach nominal significance, i.e., z-scores outside the random range $|z\text{-score}| < 1.65$. Moreover, the means of the distributions are shifted towards positive values, indicating that most disease pairs are more separated than expected by chance. Since most diseases are presumably independent from one another, this results is intuitively plausible. As expected, the difference between the two randomization models is most pronounced for negative z-scores. The full randomization model exhibits a tail of large negative z-scores which is absent in the MeSH preserving model as these z-scores correspond to pairs of related diseases and.

We further estimated the number of spurious results that may arise due to the large number (44,551) of tested disease pairs. In order to rationalize the often arbitrary choice of a significance threshold, we adopt a widely used approach from the analysis of microarray data [72, 73] to estimate the false discovery rate (FDR) directly from random simulations. For each disease pair we determine how many random simulations i yield a nominally significant result with $|z_i| > z_{\mathrm{t}}$, where the z-score threshold z_{t} is a parameter. These significant outcomes are by definition false positives, so the false positive rate f_{AB} of disease pair AB can be identified with the fraction of all $R = 1,000$ random simulations with a significant outcome:

$$f_{\mathrm{AB}}(z_{\mathrm{t}}) = \frac{\#\{|z_i| > z_{\mathrm{t}}\}}{R}\,. \tag{S11}$$

The expected overall false discovery rate across all $P = 44,551$ disease pairs can now be estimated *via*

$$\mathrm{FDR}(z_{\mathrm{t}}) = \frac{1}{P}\sum_{\{\mathrm{AB}\}} f_{\mathrm{AB}}\,, \tag{S12}$$

where $\{\mathrm{AB}\}$ denotes all combinations of diseases A and B. We find that a global FDR of 5% is reached at $z_{\mathrm{t}} \approx 2.0$ (Figure S7a; the two randomization models yield practically identical results, as they only differ in a small number of pairs). The expected number of truly significant disease-disease relations can now be estimated by adjusting the number of nominally (uncorrected) significant

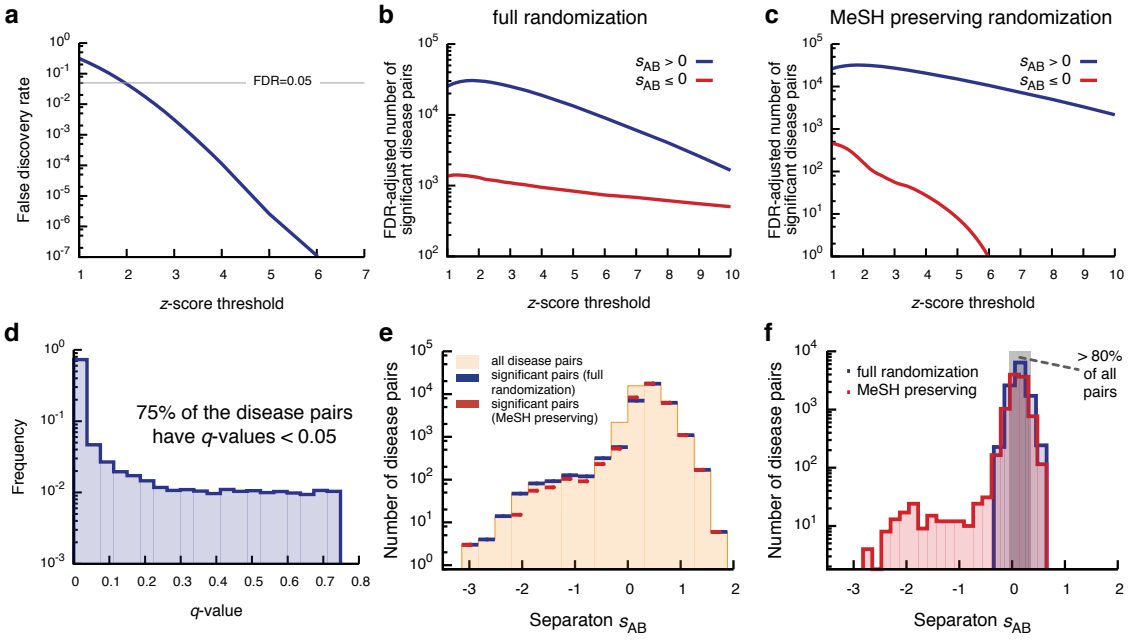


Figure S7: **Significance of network separation.** **a,** False discovery rate (FDR) *vs.* applied z-score threshold z_t (the two randomization models are indistinguishable). **b,c,** FDR-adjusted expected number of significant disease pairs with $s_{AB} > 0$ and $s_{AB} \leq 0$ *vs.* applied z-score threshold for the two randomization models. **d,** Distribution of q-values across all disease pairs. The q-value of a disease pair indicates the minimal global FDR at which it reaches significance. **e,** Distribution of s_{AB} for all disease pairs, as well as for pairs that are significant with q-value < 0.05 according to the two randomization schemes. **f,** Distribution of s_{AB} for disease pairs that are *not* significant at a global FDR of 5%. The gray area indicates the range $-0.05 \leq s_{AB} \leq 0.35$ that contains more than 80% of all such pairs for both randomization models.

pairs with (1-FDR). Figure S7b,c shows the FDR-adjusted number of significant disease pairs as a function of z_t, separately for pairs with positive and negative s_{AB}. The results confirm that indeed the network-based separation is significant for a large number of disease pairs. Comparing Figures S7b and c, we observe a smaller number of significant pairs with negative s_{AB} in the MeSH preserving randomization scheme, indicating that effects from shared genes often dominate the network-based overlap for MeSH-related diseases.

In addition to estimating the *global* FDR, the approach presented above also enables us to further quantify the level of confidence for the observed separation of each individual disease pair. For this we use the so-called q-value [73], defined as the minimal global FDR at which the particular disease pair reaches significance. Figure S7d shows the distribution of q-values for all disease pairs. In

agreement with our previous results, for small q-values we find a clear deviation from the flat distribution expected for purely random events. Indeed, 75% of all disease pairs have q-value < 0.05, again indicating that the majority of observed network relationships are non-random. We further examined the relation between the magnitude of s_{AB} and its significance. Figure S7e compares the distribution of s_{AB} of all disease pairs with the respective distributions of significant pairs only (at q-value < 0.05). For both randomization models we find that the significant pairs exhibit the full range of s_{AB} values. In Figure S7f we only show the s_{AB} distribution of disease pairs that did not reach significance. The MeSH preserving model identifies a number of disease pairs whose overlap is not more surprising than expected from their MeSH relation. Yet, overall their number is relatively low and the majority of overlapping disease pairs are found significant. In both randomization models, the vast majority of insignificant pairs have small positive s_{AB} values: more than 80% of all disease pairs whose network separation is indistinguishable from random are found in the small range $-0.05 < s_{\mathrm{AB}} < 0.35$.

In summary, we conclude that the proposed separation measure s_{AB} offers a robust quantification of the network-based relationship between diseases. As expected, we find that most disease pairs are clearly separated ($s_{\mathrm{AB}} > 0$), however, we also identified a considerable number of disease pairs with statistically significant overlap ($s_{\mathrm{AB}} < 0$).

2.3 Generalization for directed networks.

Throughout our analysis, we treated the interactome as an undirected network. Yet, the interactome does contain a number of regulatory and signaling interactions that are directed. Note that there is very little robust information available for the directionality of interactions in the current interactome data: less than 6% of the links in our network have an unambiguous direction. Yet, as illustrated in Figure S8a,b, in principle the presence of directed links can affect the extent to which local perturbations spread in the interactome. Moreover, link directionality is important for the biological interpretation of the map of a specific disease module and for designing experiments to test the predictions derived from it. In order to estimate the extent to which the presence of directed links affects our main results, we repeated our analysis on a modified directed interactome and with a generalized distance measures.

The two sources of the interactome that contain link directions are the kinase-substrate network and signaling interactions (see Section 1.1 above). We only considered the direction of a specific interaction if it was unambiguous, i.e., if the two sources reported the same direction, and

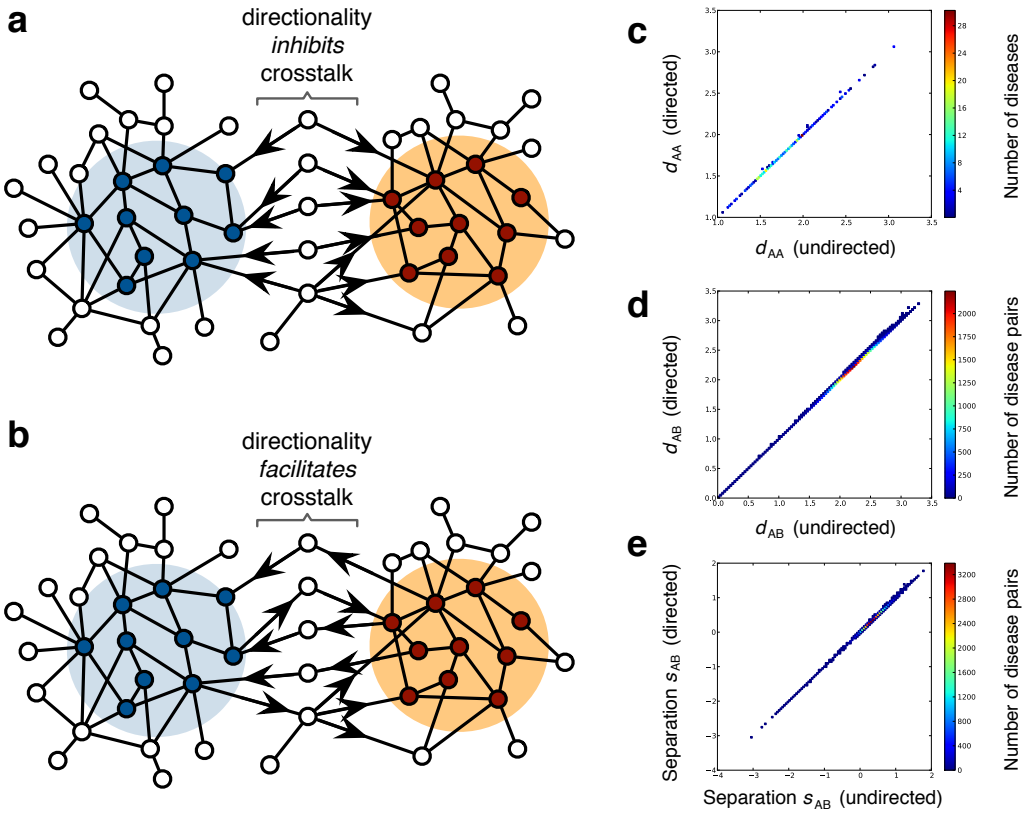


Figure S8: **a,b,** Schematic illustration of the potential impact of directed links on the separation of two disease modules. In **a**, the directed links inhibit the spread of a perturbation from either module to the other. In contrast, in **b** they facilitate interactions between the modules. **c-e,** Comparison of the network-based quantities d_{AA} (**c**), d_{AB} (**d**) and s_{AB} (**e**) between the undirected interactome and a modified version with directed links. The fact that the data points lie along the diagonal indicates that the direction does not affect the measured s_{AB} values.

if there was no simultaneous undirected interaction for the same protein pair, for example, from the binary y2h dataset. In total, we find 7,863 directed links (3,785 signaling and 4,078 kinase-substrate interactions), representing 5.6% of all interactions in the interactome. We assume that all other interactions act in both directions and consequently replaced each undirected link by two directed links pointing in opposite directions, resulting in an interactome in which all interactions are directed.

On this directed interactome, we recalculated our distance measures introduced above by replacing the undirected network distances with the respective distances along directed paths [64, 65]. Figure S8c-e compares the values of d_{AA}, d_{AB} and s_{AB} for all diseases and disease pairs in the directed and the undirected interactome. We find that the results are not affected by the directionality of the links, which may be due to the low fraction of directed links in our data. The link direction may play an increasingly important role as more complete and accurate directed data become available. Yet, directed links do not represent a methodological roadblock to our approach.

3 Comparison with Gene-based Overlap Measures

In this section we compare the introduced network-based overlap of diseases with overlap measures that are solely based on shared genes.

3.1 Gene-set overlap

We use two measures to quantify the overlap between two gene sets A and B:

$$\text{overlap coefficient } C = \frac{|\, A \cap B \,|}{\min(|\, A \,|, |\, B \,|)} \quad , \tag{S13}$$

$$\text{Jaccard-index } J = \frac{|\, A \cap B \,|}{|\, A \cup B \,|} \quad . \tag{S14}$$

Their values lie in the range $[0, 1]$ with $J, C = 0$ indicating no common genes in both cases. A Jaccard-index $J = 1$ indicates two identical gene sets, whereas the overlap coefficient $C = 1$ when one set is a complete subset of the other. Figures S9a,b show the distribution of C and J for all 44,551 considered disease pairs. The overlap is relatively low for most disease pairs, the majority (59%) do not share any genes. For a statistical evaluation of the observed overlaps we use a basic hypergeometric model with the null hypothesis that disease associated genes are randomly drawn from the space of all N genes in the network. The overlap expected for two gene sets A and B is then given by

$$c_{\mathrm{rand}} = \frac{|\, A \,| \times |\, B \,|}{N} \,. \tag{S15}$$

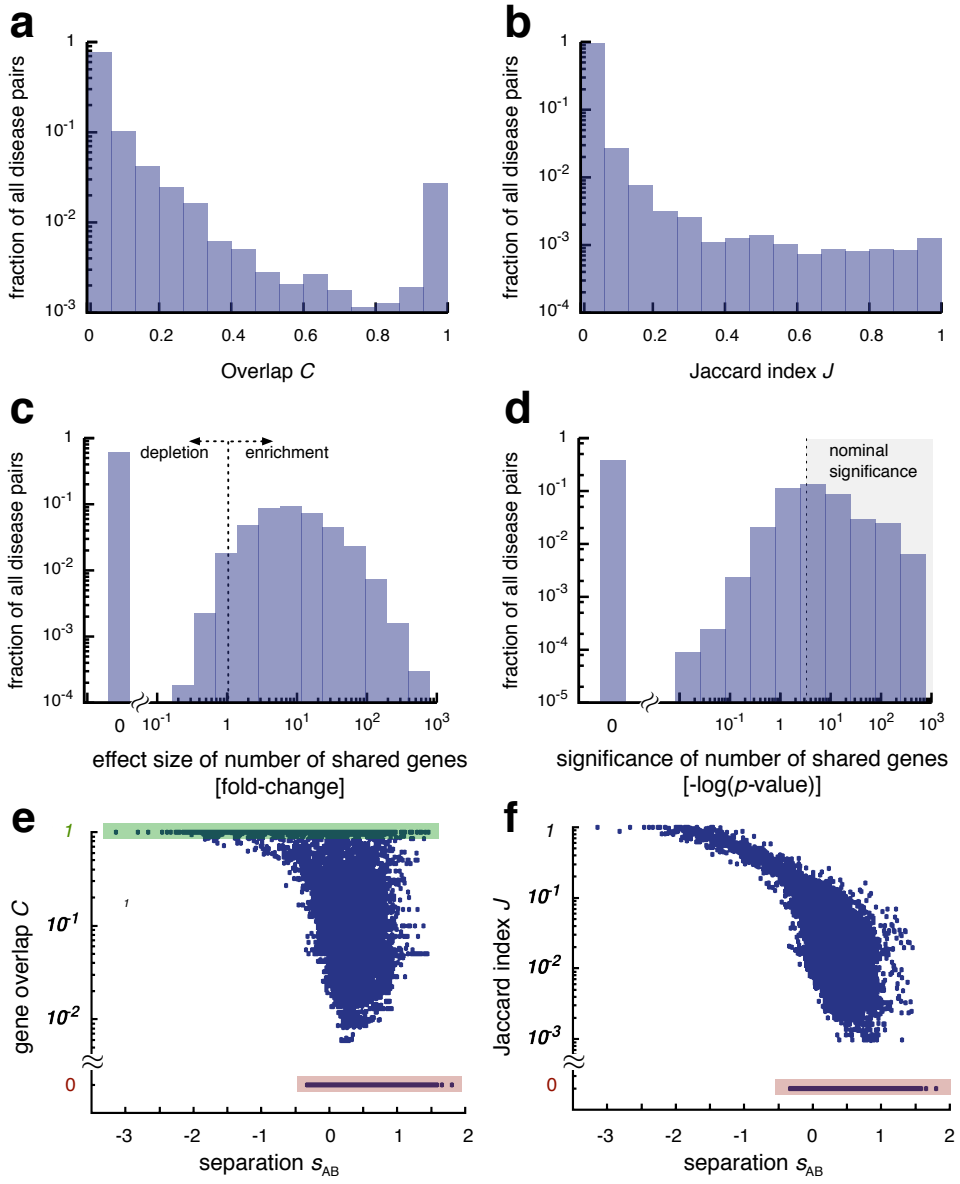


Figure S9: **Comparison of network- and gene-based overlap.** **a,b** Distributions of C and J across all disease pairs. **c,** Distribution of the fold-change of the number of shared genes compared to random expectation across all disease pairs; fc < 1 indicates depletion, i.e., fewer common genes than expected, whereas fc > 1 indicates enrichment. **d,** Distribution of the statistical significance of the observed number of shared genes. Nominal (i.e., uncorrected) significance p-value < 0.05 is indicated by the shaded area. **e,f** Phase-diagrams showing the gene-overlap measures C and J *vs.* the network-based separation s_{AB} for all 44,551 considered disease pairs. The green area highlights the 1,101 pairs with $C = 1$, i.e., for which the genes of disease A are a complete subset of the genes of disease B. The red are highlights disease pairs without common genes, representing 59% of all disease pairs.

For every observed overlap $c_{\text{obs}} = | A \cap B |$ we then determine the fold-change

$$\text{fc} = \frac{c_{\text{obs}}}{c_{\text{rand}}} \tag{S16}$$

and separately the p-values for enrichment and depletion, i.e., for a surprisingly high or low overlap, from the full hypergeometric distribution (Figures S9c,d). Note that these two tests are equivalent to the two one-tailed Fisher exact tests. We find that the overlap of 98% of all disease pairs with at least one common gene is larger than expected (fc > 1). For 60% of these pairs the enrichment reaches nominal significance (p-value < 0.05), 23% are significant even after applying the most conservative Bonferroni correction. Depletion, on the other hand, is virtually absent, i.e., there are no disease pairs for which the observed overlap is significantly smaller than expected (only 0.02% of all pairs with fc < 1 reach nominal significance, none after correction).

3.2 Network-based overlap and gene-set overlap

In the previous section we found that the basic statistical evaluation of the gene-based overlap suggests significant overlap for most disease pairs that share genes. Diseases without shared genes, on the other hand, cannot be further differentiated. We now explore the relation between the gene-based overlap measures and the network-based separation introduced in this work. Figure S9e,f shows the gene-based overlap measures C and J *vs.* s_{AB} for all 44,551 considered disease pairs. For disease pairs with common genes the two measures generally correlate. However, 59% of all disease pairs do not share any associated genes ($C = J = 0$), hence the relationships of more than half of all possible disease pairs cannot be quantified using gene-based measures. In contrast, our network-based approach reveals that these disease pairs exhibit a range of different relationships, from overlapping disease modules ($s_{\text{AB}} < 0$) to strictly separated disease pairs ($s_{\text{AB}} > 0$). On the other side of the spectrum, we also find a number of diseases with considerable gene overlap, but insignificant network similarity. Even disease pairs for which the gene set of one disease is a complete subset of the other ($C = 1$; 1,101 disease pairs) can exhibit the full range of network-based relations, from overlapping to separated disease modules (Figure S9e).

3.3 Additional control sets for the network-based disease similarity

In the main text we compare the biological similarity of disease pairs with and without network-based overlap for all considered disease pairs (Figure 4b-g) and disease pairs without common genes (Figure 4h-m). To further substantiate our finding that the network-based separation measure is predictive for biological similarity we repeated the analysis on two additional sets of disease pairs,

see Figure S10. (*i*) As indicated above, our disease corpus includes by construction diseases at different levels of the MeSH hierarchy, such that one disease is a more specific variant of the other. (compare with Figure S2). There are relatively few of these disease pairs (1,101; representing 2.5% of all pairs), yet in order to confirm that our results are not distorted by disease pairs, we have repeated our analysis removing all respective pairs. Figure S10c shows that the overall results prevail. (*ii*) We also considered the opposite situation, where the genes associated with one disease are a complete subset of the genes of the other ($C = 1$). Consequently, the diseases exhibit very high levels of biological similarity. Yet, almost half (47%) of the respective pairs have $s_{\mathrm{AB}} > 0$, i.e., from a network perspective they are separated. We find that even for the disease pairs with high disease gene overlap, the network separation reveals significant differences in biological similarities (Figure S10d). These pairs mostly represent a broad disease category, like *nutritional and metabolic diseases*, and a specific variant of it, like *obesity*. Their separation indicates that the proteins associated with the specific variant reside in a different neighborhood of the interactome – indicating that the variant may lead to quite different disease mechanisms than the broad disease class. These cases, therefore, reflect shortcomings of traditional disease classification that is based on clinical manifestations rather than molecular origins. Indeed, the molecular mechanisms of obesity are expected to be more closely related to disorders of the hypothalamus and its satiety center and perturbations in the endocrine system through gut and brain hormones that affect hunger and satiety rather than nutritional diseases, thereby justifying the separation of obesity from the latter.

In summary, we find in all considered controls that the network-based measure is highly predictive for biomedical similarity, thereby offering additional insights into disease pairs for which gene-based overlap measures cannot offer further discrimination.

4 Comparison with Unbiased Datasets

Both the interactome data, as well as the disease associations are prone to investigative biases. We therefore systematically explore how our results are affected when only high-throughput data, or combinations of high-throughput and literature curated data are used.

Using only yeast-two-hybrid (y2h) interaction data. Figure S11a shows the observed module sizes for all diseases as a function of the total number of disease genes when only high-throughput interactions from y2h (see Sec. 1.1.2) are used instead of the full interactome. In total, 34 out of 299 diseases show a statistically significant module. According to our results from percolation theory

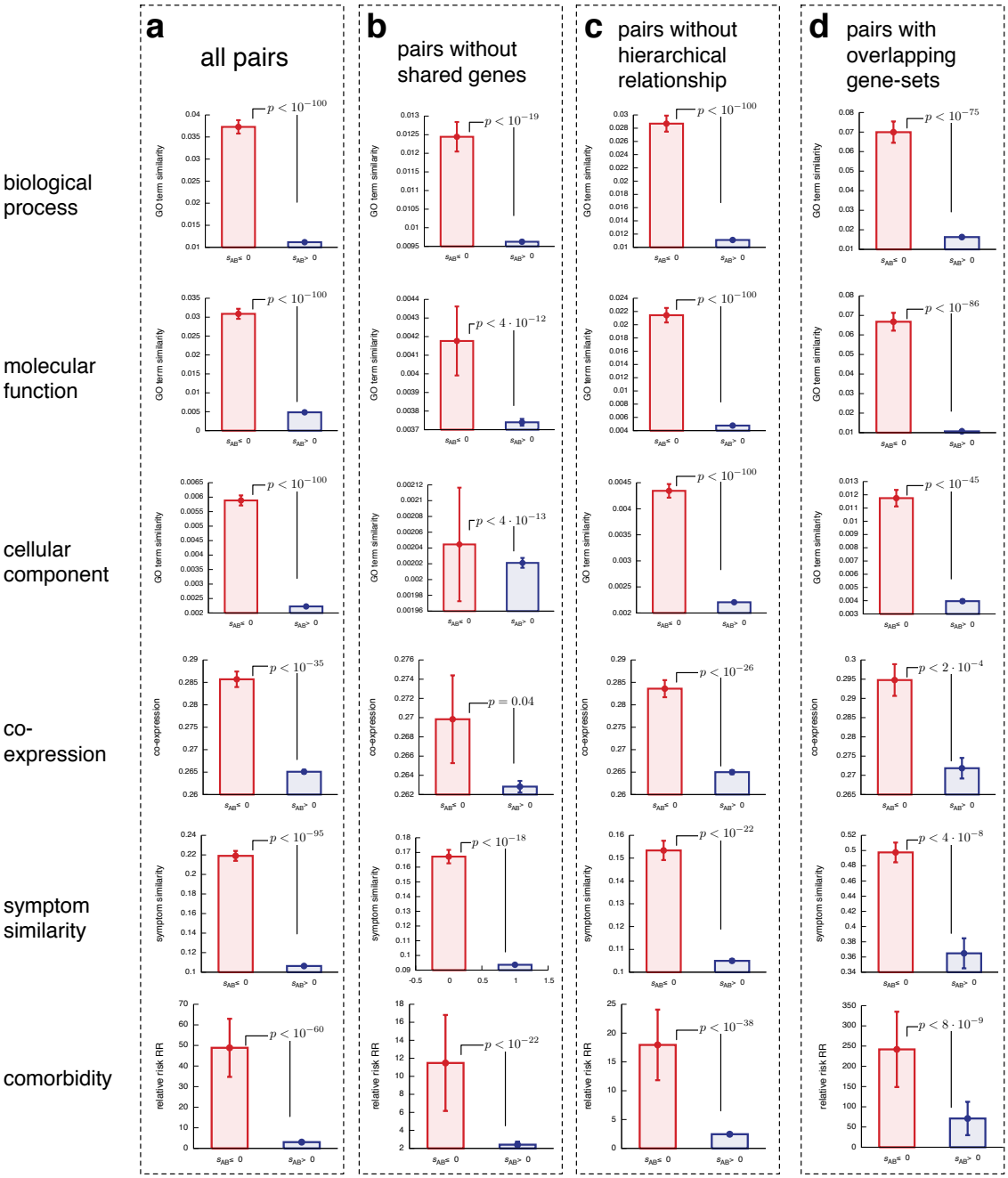


Figure S10: **Additional control sets for the network-based disease similarity.** Comparison of the biological similarity of overlapping ($s_{\mathrm{AB}} < 0$) and non-overlapping disease pairs ($s_{\mathrm{AB}} > 0$) for four different sets of disease-pair selections. Error bars show the standard error of the mean, p-values are computed using a Mann-Whitney U test. **a** and **b** show the same results as in Figure 4b-m in the main text and are repeated here for comparison: **a,** all disease pairs. **b,** only disease pairs without common genes. **c,** Only disease pairs without hierarchical MeSH relation. For example, *prostatic neoplasms* vs. *neoplasms* has been excluded (compare with Figure S2a). **d)** Only disease pairs, where the gene set of one disease is a complete subset of the other, i.e., $C = 0$ (compare with Figure S9).

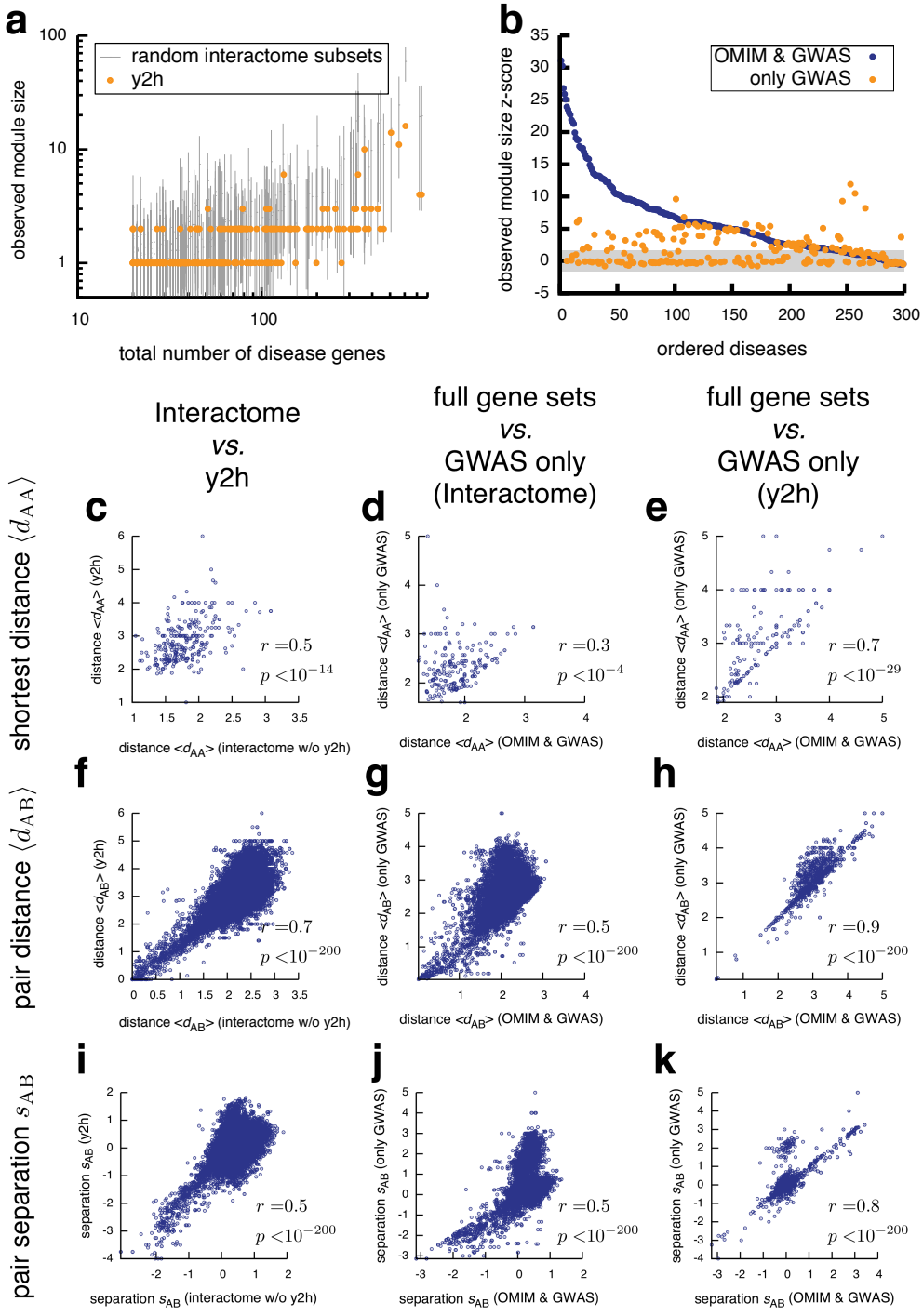


Figure S11: **Localization in unbiased datasets. a,** Size of the observed modules in the yeast-two-hybrid network (y2h). Bars indicate the module sizes obtained from random removal of links from the full interactome. **b,** Significance of the observed modules when only GWAS disease associations are used. **c-k,** Correlation of network localization measures between the full datasets and unbiased high-throughput data only. Note that in c,f,i all interactions of the y2h data have been removed from the interactome in order eliminate possible confounding factors in the correlation analysis.

this is expected, given that the y2h dataset contains only 15,937 interactions between 4,612 proteins. To further confirm this, we also constructed 100,000 random subnets from the full interactome with the same number of proteins and links as in the y2h dataset and measured the module sizes. We find that the module sizes as measured in the y2h network lie within the expected range (see bars in Figure S11a). We also computed the shortest distances for all diseases within the y2h interactome, finding that 62 diseases have statistically significantly shorter distances than expected by chance.

Using only GWAS disease associations. Figure S11b shows the module sizes within the full interactome when only disease associations from GWAS are used. We find that 96 of the 218 diseases that have gene associations from GWAS show a statistically significant module size, a surprisingly high number given that there are considerably fewer disease genes as compared to the full set that includes associations from OMIM.

Comparing network localization using different datasources. In Figure S11c-k we analyse the values of the mean shortest distance $\langle d_{\mathrm{AA}} \rangle$ of all diseases (c,d,e), as well as the pairwise distances $\langle d_{\mathrm{AB}} \rangle$ and separation s_{AB} (i,j,k) for different combinations of high-throughput data only, compared to data that includes literature curated data. Figure S11c,f,i shows strong and highly significant correlations between the respective values of $\langle d_{\mathrm{AA}} \rangle$, $\langle d_{\mathrm{AB}} \rangle$ and s_{AB} within the interactome and those within the y2h network. Note that we have removed all links contained in the y2h data from the interactome, even those for which additional literature evidence exists, thereby ensuring that the observed correlations are not an artefact due to common links. Equally strong correlations are found when comparing the values obtained using the full gene sets to those obtained using GWAS gene associations only, both in the full interactome (Figure S11d,g,j) and the y2h network (Figure S11e,h,k).

Taken together, we find that within the limitations imposed by the sparser data, the localization of disease proteins in the human interactome systematically persists also in unbiased high-throughput data and that the values of $\langle d_{\mathrm{AA}} \rangle$, $\langle d_{\mathrm{AB}} \rangle$ and s_{AB} correlate strongly between different combinations of datasources.

Biological similarity for overlapping and separated diseases. Since the topological properties of diseases strongly correlate between full and restricted unbiased datasets, the main findings shown in Figures 3 and 4 of the main text are expected to be conserved. In order to test this explicitly, we repeated the analysis of our main results concerning the predictive power of the network-based sep-

aration for biological similarity in Figure 4b-m of the main text. The results provided in Figure S12 indeed confirm that our main finding can be reproduced even in the much sparser y2h network.

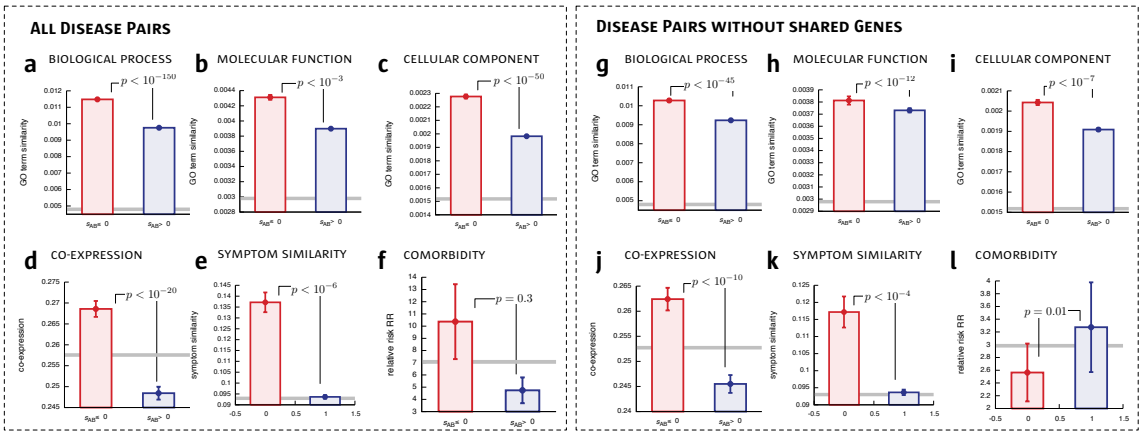


Figure S12: **Network separation and biological similarity in the unbiased y2h network.** Biological similarity shown separately for the predicted overlapping and non-overlapping disease pairs, compare with Figure 4b-m of the main text . Error bars indicate the standard error of the mean. Gray lines indicate random expectation, either for random protein pairs (a-d,g-j) or for random disease pairs (e,f,k,l), p-values denote the significance of the difference of the means according to a Mann-Whitney U test. **a-f** show the biological similarity for all considered disease pairs, **g-l** for the subset of pairs that do not share genes (control set).

5 False Positive Links and Network Localization

Current interactome maps are expected to contain a considerable number of false positive interactions. We therefore systematically explored the extent to which such links could lead to a spurious clustering of disease proteins. For both the interactome, as well as the unbiased hight-throughput y2h dataset we artificially increased the false positive rate by introducing random connections into the network. We use two mechanisms: (i) Completely randomly scattered links by repeatedly choosing two proteins from the network at random and connecting them. (ii) Choosing two proteins with a probability proportional to their degree in the original network, thereby mimicking the effects that may arise from highly-studied proteins that may also have a higher local false-positive rate of interactions. For varying levels of introduced random links we then measured the significance of the largest connected component for one well localized disease (*multiple sclerosis* for the interactome and *genetic skin diseases* for the y2h network), as well as one non-localized disease

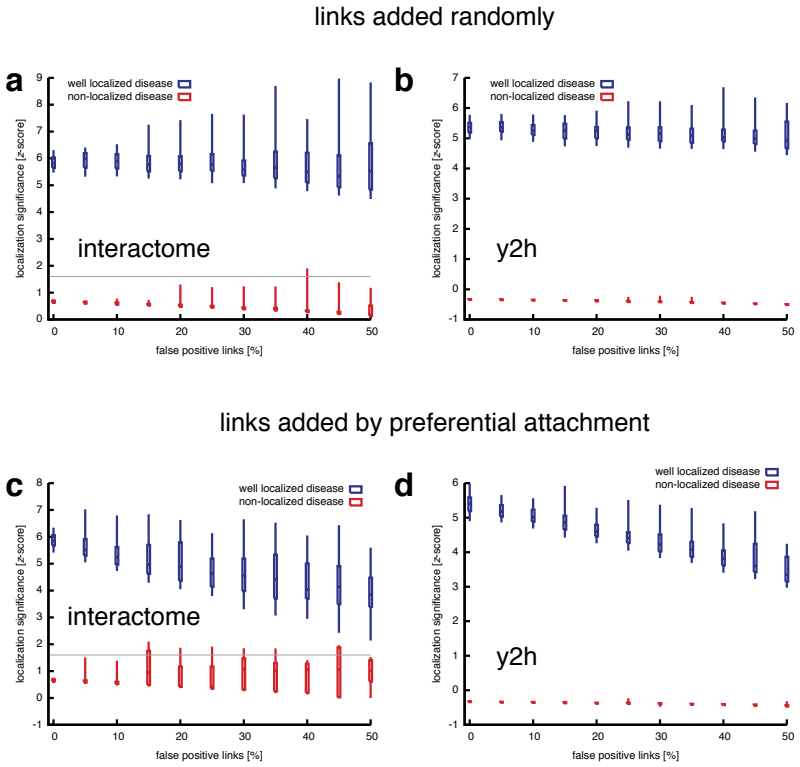


Figure S13: **Observable module size and false positive interactions.** To evaluate the impact of false positive interactions on a well-localized and a non-localized disease we compute the significance of the size of the respective observable disease modules for different fractions of added random links. Gray lines indicate the significance level $z - \text{score} = 1.6$. **a,b,** Completely randomly added links for the interactome (a) and the high-throughput y2h network (b). **c,d,** Random links are added according to the degree of the proteins. We observe that in general false positive interaction only increase the noise-level, but do not artificially inflate the disease modules.

(*cerebellar ataxia* for the interactome and *multiple sclerosis* for the y2h network) using random simulations as described above in Section 2. Figure S13 shows the results from 1,000 randomly inflated networks with 1,000 randomized disease protein configurations for each false positive rate. As expected, we find that random links generally do not increase the localization of neither the well localized, nor the non-localized diseases. To the contrary, in the case of the degree adapted randomization scheme the degree of localization is diminished as the fraction of false positive links increases (Figure S13c,d). We conclude that our initial observation of well-localized disease modules is not a consequence of false positive interactions.

6 Identifiability of Disease Modules

In the following we quantify the fragmentation of disease modules in incomplete interactome data. We derive several exact results that are independent of many network parameters, and hence, offer reliable predictions pertaining to the identifiability of disease modules.

The current interactome can be viewed as an incomplete version of the underlying complete network, in which both links and nodes are missing [74–77]. The y2h dataset [12, 55], for example, results from a systematic screen of all pairwise combinations between a set of proteins comprising $\sim 66\%$ of all known proteins, i.e., $\sim 33\%$ of the nodes are missing, together with all their links. The number of missing links is further increased by limitations of the experimental assay, whose overall sensitivity is estimated to be $\sim 20\%$. The effects of removing nodes and/or links from a network have been studied extensively in the framework of percolation theory [78–81]. In the most general setting, it has been found that as long as a certain critical fraction of all N nodes (or M links) is present in the network, it exhibits a *giant component*, i.e., a connected subgraph, which contains a number of nodes comparable to the size of the complete network (Figure S14a). The fraction of nodes in that giant component S can be computed using the formalism of generating functions.

Let $P(k)$ be the degree distribution of a network and $Q(k)$ the probability that the node at the end of a randomly chosen link has degree k. The generating functions of these two quantities are

$$G_0(x) = \sum_k P(k) x^k \tag{S17}$$

$$G_1(x) = \sum_k Q(k) x^k \,. \tag{S18}$$

The detailed shape of the curve in Figure S14a depends on the network structure and the details of the link and node removal. For unbiased incompleteness, i.e., when all nodes (links) have the same probability p of being present in the network, the size of the giant component S can be computed by solving

$$S = \begin{cases} p[1 - G_0(u)] & \text{[node percolation]} \\ 1 - G_0(u) & \text{[link percolation]} \end{cases} \tag{S19}$$

$$u = 1 - p + p\, G_1(u) \,. \tag{S20}$$

The *percolation threshold*, i.e., the critical completeness p_c at which S appears and the previously disconnected small fragments merge to a giant component is given exactly by

$$p_c = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} \,, \tag{S21}$$

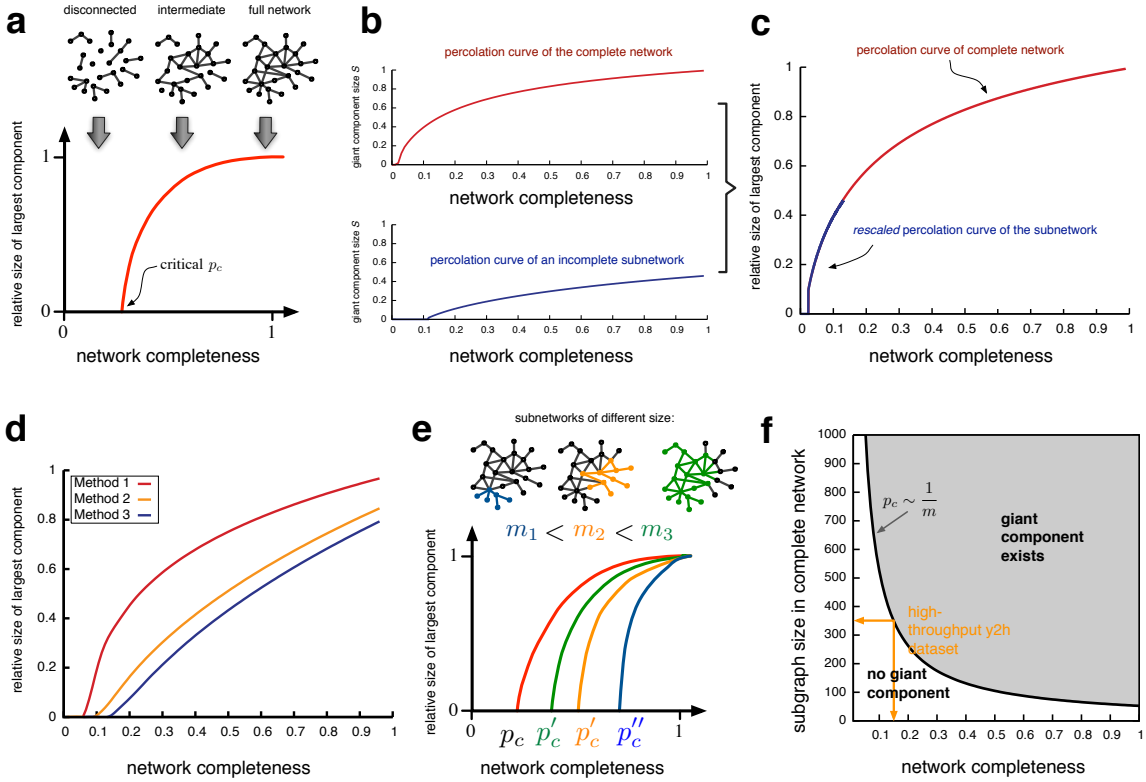


Figure S14: **The Identifiability of Disease Modules. a,** The relative size of the giant component in function of network completeness. Generally, the completeness is given by the product of the fractions of all nodes p and all links q that are present in the incomplete network. At $p\,q = 1$ all nodes and links are present and the network has a single connected component. As more and more nodes/links are removed, the size of the giant component shrinks until it vanishes at the critical completeness p_c. **b,** Percolation curves (link percolation) for the interactome (top) and a random subset of it (bottom), obtained by removing 33% of the nodes and 80% of the links. **c,** The curves from **b** collapse when the curve of the incomplete subset is rescaled according to Eq (S23). **d,** Comparison of the percolation curves (link percolation) for three connected subgraphs of size $m = 100$ in the interactome that were constructed using different methods. **e,** Schematic illustration of the percolation curve in **a** when subgraphs of size m are considered instead of the whole network. Generally, the percolation threshold $p_c(m)$ will be larger for smaller subgraphs, i.e., smaller subgraphs require a higher completeness in order to be observable. **f**, Phase diagram showing the minimal completeness needed, such that a module of a given size in the complete network still exhibits a giant component in the incomplete network.

where $\langle k \rangle$ and $\langle k^2 \rangle$ denote the first and second moment of the degree distribution $P(k)$. Therefore $P(k)$ is sufficient to compute the critical completeness, as well as the expected size of the giant componen at any given level of network completeness above the threshold. Note that the percolation threshold is the same for node and link removal.

While the general framework introduced above predicts the percolation threshold, it requires knowledge of the complete network, information that is not currently available for the interactome. We can circumvent this problem by recognizing that *the percolation curve of a complete network can be partially reconstructed from a sample of the network*, provided that the sample is uniform. To demonstrate this principle, let us assume that the current interactome is complete. Figure S14b (top panel) shows the corresponding percolation curve $S^{\rm full}(p)$ (link percolation), predicted by Eqs. (S17)–(S21). To simulate the effects of the network incompleteness introduced above (66% of all proteins screened with a sensitivity of 20%), we generated a second, smaller network by randomly removing 33% of the nodes and subsequently 80% of the remaining links. The percolation curve $S^*(p)$ for this pruned network is shown in Figure S14b (bottom panel). By construction, this second network represents a version of the original one at the completeness level $p = p^*\, q^*$ with

$$p^*\, q^* = (\text{fraction of screened proteins})\ p^* \times (\text{sensitivity to detect a link})\ q^*\,. \tag{S22}$$

It follows that we can obtain a fraction of the original –and in the case of the true interactome unknown– percolation curve by rescaling the curve measured of the incomplete network as

$$S^{\rm full}(p) = S^*(p\, p^* q^*)\,. \tag{S23}$$

Figure S14c shows the collapse of the two curves in Figure S14b after rescaling according to Eq. (S23). We see that $S^{\rm full}(p)$ can be recovered up to $p^*\, q^*$. In particular, we can obtain the critical percolation threshold $p_c^{\rm full}$ of the unknown full network from measuring $p_c^{\rm y2h}$ of the available incomplete network:

$$p_c^{\rm full} = p_c^{\rm y2h} p^*\, q^*\,. \tag{S24}$$

This relation is exact for arbitrary networks and can be directly applied to the y2h dataset [12, 55], which, with good approximation, represents a uniform subset of the corresponding full interactome: For an unbiased fraction p^* of all proteins, all pairwise interactions have been tested, so the present fraction of all real interactions corresponds to the sensitivity q^* of the experimental protocol.

Thus far we viewed the current interactome as a subset of the unknown complete interactome. Next, we show that we can use the formalism introduced above to determine properties of the

connected subgraphs, allowing us to explore the integrity of the disease modules. In general, the observability of a disease module depends on the structure of the complete network, its incompleteness, and in particular on the structural details of the subgraph itself. Consider, for example, the percolation curves for three connected modules in the interactome (Figure S14d). The curves were obtained by first constructing connected subgraphs, then measuring their degree distributions $P_{\rm sub}(k)$, and finally applying the formalism from Eqs. (S17)–(S21). The three modules have the same size $m = 100$, but were constructed using different methods that result in slightly different topological properties: (*i*) Starting from a randomly chosen node, we iteratively added randomly chosen nodes from the neighborhood of the current cluster. (*ii*) Starting from a node at the end of a randomly chosen edge, we iteratively added nodes reached by following a randomly chosen edge that leads out of the current cluster. (*iii*) We placed random nodes on the empty network until a giant component of size m emerges. Comparing the three curves in Figure S14d, we find that the modules obtained by the different methods exhibit different percolation thresholds p_c. The threshold represents the critical level of incompleteness at which the subgraphs disintegrate into isolated fragments and are no longer observable as being connected in the global network. The module generated by method (*iii*) exhibits the highest p_c, i.e., the highest required completeness for observability. This is expected, given that the method is a percolative process itself and should, therefore, result in modules that are close to the percolation threshold.

We have seen that the exact percolation properties of specific submodules require a detailed knowledge of their properties in the complete network that is generally not available for disease modules. Yet, we can use the results above to estimate the *minimal completeness* at which a module of a given size m can be observed. We start by considering the ensemble of connected subgraphs obtained by random selection of nodes as in method (*iii*). By construction, these subgraphs exhibit a high threshold p_c that provides an upper limit for denser modules with lower thresholds (compare with Figure S14d). The threshold p_c, in turn, is bound from above by the threshold of the subgraphs left from the full network at node completeness $p = m/N$. We have seen above that their threshold can be obtained from Eq. (S24), so by using $m/N = p^* q^*$, we finally obtain

$$p_c(m) = \frac{N p_c^{\rm full}}{m} \tag{S25}$$

as an upper bound for the *minimal completeness* $p_c(m)$ at which we expect to be able to observe the remainders of a module of size m. Note that for brevity we referred to this quantity as p_c^m in the main text. It follows from (S22)–(S25) that for incomplete networks which are a uniform subset of the complete network, a detailed knowledge of the complete network is *not* required to determine $p_c(m)$. Instead, it is sufficient to know only the *level* of incompleteness, as $p_c^{\rm full}$ can be obtained from

p_c^{y2h}, which can be measured directly. Figure S14d illustrates schematically how the percolation curve changes for subgraphs of different sizes m. Eq. (S25) indicates that the percolation threshold increases with decreasing subgraph size, i.e., smaller subgraphs require a higher network completeness in order to exhibit a giant component. Figure S14e shows the minimal subgraph size for which we expect to find a remaining connected component for a given level of network completeness as derived from Eqs. (S17)–(S25) using the y2h network as input. The yellow arrow indicates the estimated values for the current dataset. We find that the coverage of the current y2h dataset is still too small to observe significant clustering for the given number of disease associated genes. We expect, however, that this will improve significantly, once the ongoing efforts to screen a yet larger number of proteins are completed.

The $1/m$ scaling of the percolation threshold in Eq. (S25) is valid for arbitrary degree distributions. To offer a more intuitive understanding of this general result, in the following we present an alternative derivation for the special case of Erdős-Rényi random graphs:[3] Consider a random graph with m nodes, in which every possible link is present with probability p. The expected number of links is then given by

$$M = p\frac{1}{2}m(m-1). \tag{S26}$$

According to the famous results derived in [82], the critical probability at which the giant component emerges in a random graph is given by

$$p_c^{\mathrm{ER}} = \frac{1}{\langle k\rangle}\,. \tag{S27}$$

Eq. S27 implies that random graphs exhibit a giant component for $\langle k\rangle \geq 1$, i.e., when each node has one neighbor on average. With Eq. (S26), the mean degree can be written as

$$\langle k\rangle = \frac{2M}{m} = (m-1)p \sim m\,, \tag{S28}$$

substituting in Eq. (S27) finally yields

$$p_c^{\mathrm{ER}} \sim \frac{1}{m}\,, \tag{S29}$$

which recaptures the scaling from Eq. (S25).

In summary, we have shown that the percolation formalism can be used to quantify the detectability of modules in the interactome. Our first key finding is that a detailed knowledge of the complete network is *not* required to determine its percolation threshold. Given a uniform subset of

[3] We thank an anonymous Referee for bringing this to our attention.

the complete network, it is sufficient to know only the *level* of incompleteness. This result is exact for arbitrary degree distributions. Our second key result is the inverse relationship, $p_c(m) \sim 1/m$, between the size of a module and the minimal network completeness for its observability. This relationship is exact for Erdős-Rényi graphs and provides a lower bound for the general case of arbitrary networks. We expect that these general results will also find applications beyond the presented case of disease modules in the interactome.

7 Comparison of Disease Modules and Network Communities

In principle, one can distinguish between (*i*) topological, (*ii*) functional and (*iii*) disease modules [6]. Topological modules are generally defined as locally dense network areas [83]. Functional modules are sets of proteins that are associated with a specific biological function. Disease modules represent the set of proteins whose perturbation is associated with a particular pathophenotype. From a network science perspective, one may ask to what extent disease modules and topological modules coincide. We evaluated this hypothesis using three representative, methodologically distinct community detection algorithms: (*i*) an algorithm that is based on link-similarities (LC) [84], (*ii*) the *Louvain* method (LM), which maximizes a global modularity function [85], and (*iii*) the flow-based Markov Cluster algorithm (MCL) [86]. Each of these methods identifies a large number of communities within the interactome (LC: 61,647; LM: 163; MCL: 2,029), containing up to several thousands of proteins. In order to evaluate whether some of these communities may be candidates for specific diseases modules, we determined their enrichment with genes associated with any of the 299 diseases. We find that only between 1%-5% of the communities detected by the different methods show nominal significance (p-value < 0.05, Fisher's exact test). However, even these few associations are most likely spurious, as the respective communities are small and mostly contain only a single disease genes. Not a single identified enriched community included the full observable module (i.e., the largest connected component) of any disease. As we have shown in the main text, disease modules are indeed not particularly densely interconnected and therefore do not represent topological modules in the traditional sense established in network science.

Similarly, existing topological community detection methods are also not adequate to quantify the overlap of diseases modules as they define overlap simply based on common nodes. The concept of overlapping modules introduced in our manuscript is fundamentally different as it allows for overlapping modules even when no proteins are shared.

8 Disease Space Layout Algorithm

In order to display the network-based relationships between diseases visually, we introduce a three-dimensional *disease space*, in which diseases are represented by spheres with diameter $\langle d_s \rangle$. The spatial distance r_{AB} between two diseases approximates their network-based distance $\langle d_{\mathrm{AB}} \rangle$. An exact mapping is not possible in three dimensions, since the network distances of all 44,551 disease pairs impose many conflicting constraints which cannot be fully resolved. Related problems of placing points into an n-dimensional space according to a given distance matrix have been studied extensively in the field of *multidimensional scaling*. Here, we use the following algorithm aiming to match spatial and network-based distances as closely as possible (see Figure S15): We start with a

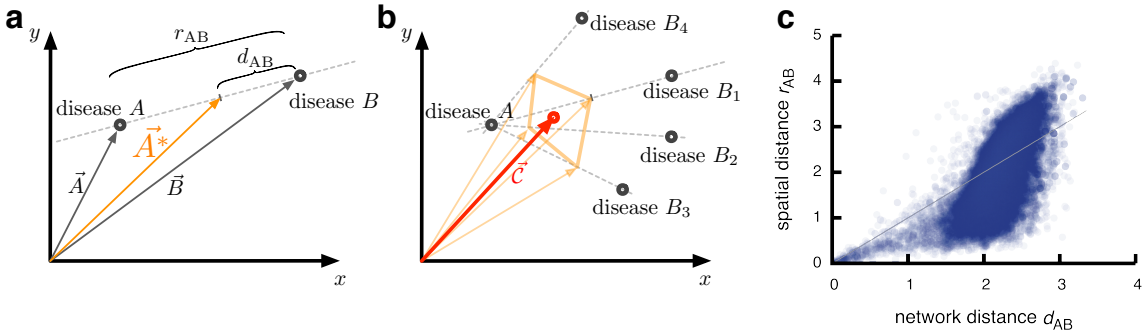


Figure S15: **Disease Space layout algorithm.** **a,** Illustration of the calculation of the optimal position $\vec{A}^*$ of disease A with respect to disease B. **b,** The new position of disease A is given by the center of mass $\vec{\mathcal{C}}$ of all pairwise optimal postions $\vec{A}_i^*$ with respect to all other diseases B_i. **c,** Comparison of the final spatial distance r_{AB} versus network distance d_{AB} for all disease pairs.

random initial placement of all diseases. Next, we individually update the postion of each disease according to the following procedure: Starting from the position of disease A, given by the vector $\vec{A}$, we calculate for all other diseases B_i a new position $\vec{A}_i^*$ along the lines that connect $\vec{A}$ and $\vec{B}$, such that the spatial distance r_{AB} between $\vec{A}_i^*$ and $\vec{B}_i$ matches exactly their network distance d_{AB} (Figure S15a):

$$\vec{A}_i^* = \vec{B} + \frac{d_{\mathrm{AB}}}{r_{\mathrm{AB}}} \left(\vec{A} - \vec{B}_i \right) . \tag{S30}$$

Disease A will then be moved to the center of mass $\vec{\mathcal{C}}$ of all pairwise optimal positions $\vec{A^*}_i$ (Figure S15b):

$$\vec{\mathcal{C}} = \frac{1}{n} \sum_{i=1}^{n} \vec{A^*}_i \, . \tag{S31}$$

In this fashion we repeatedly update the position of all diseases until they converge. While the exact final coordinates do depend on the random initial configuration, we find that the total mismatch

between spatial and network-based distance for all disease pairs $\sum_{\text{pairs}} |d_{\text{AB}} - r_{\text{AB}}|$ is rather robust and varies very little between different random realizations. Figure S15c shows d_{AB} and r_{AB} for all disease pairs for the coordinates that we used to layout the spheres in Figure 4a,h in the main text. We find the spatial and network-based distance to be strongly correlated.

9 Discussion of selected disease pairs

As discussed in the main text, we identified several topologically overlapping disease pairs that lack known pathobiological relationships. Here we briefly discuss twelve additional examples, as summarized in Table S1.

disease A	disease B	N_A	N_B	$N_A \cap N_B$	common genes	L_{AB}	s_{AB}	RR
asthma	celiac disease	37	36	3	HLA-DQA1, IL1RL1, IL18R1	7	-0.13	6.1
glomerulonephritis	liver cirrhosis, biliary	18	23	2	HLA-DQB1, HLA-DQA2	4	-0.05	1.6
psoriasis	vasculitis	54	15	9	IER3, DDR1, MICA, IL23R, HLA-B, HLA-C, PSORS1C2, C6orf15, CCHCR1	11	-0.35	1.8
graves disease	vasculitis	13	15	2	C6orf15, MICA	0	-0.30	1.9
lymphoma	myocardial infarction	24	13	0	–	3	-0.24	2.1
glioma	myocardial infarction	17	13	0	–	3	-0.19	6.3
metabolic bone diseases	myocardial infarction	27	13	0	–	1	-0.15	1.4
glioma	gout	17	13	0	–	2	-0.14	2.4
liver cirrhosis	spondylitis, ankylosing	24	12	0	–	3	-0.20	1.0
leukemia, lymphocytic, chronic, b-cell	motor neuron disease	15	31	0	–	6	-0.14	1.2
albuminuria	asthma	15	37	0	–	2	-0.13	1.2
myeloproliferative disorders	proteinuria	19	15	0	–	0	-0.14	1.9

Table S1: **Basic characteristics of chosen disease pairs with overlapping modules within the interactome.** N_A, N_B: number of genes associated with the diseases A and B, respectively; $N_A \cap N_B$: number of common associated genes; L_{AB}: number of direct interactions between proteins associated with A and B; s_{AB}: network-based separation; RR: comorbidity as measured by the relative risk.

Glomerulonephritis & biliary cirrhosis. *Glomerulonephritis*, an inflammatory renal disease, and *biliary cirrhosis*, an autoimmune disease of the liver, are associated with 26 and 32 genes, respectively, and show moderate localization in the same network neighborhood ($s_{\text{AB}} = -0.05$). They

share two GWAS genes (HLA-DQA2, HLA-DQB1). Although the overall gene annotations are not significantly similar, they do share significant pathways and are highly comorbid ($RR = 1.6$). It has previously been observed that *primary biliary cirrhosis* is associated with *membranous glomerulonephritis* in some patients [87].

Psoriasis & vasculitis. *Psoriasis*, an immune-mediated skin disease, and *vasculitis*, an inflammatory disease affecting blood vessels, have 77 and 36 known gene associations, respectively. Both are well localized and overlap in the interactome ($s_{\mathrm{AB}} = -0.35$). This overlap also reflects the large number (17) shared genes, all of which are derived from GWAS. Overall the gene annotations are similar and the genes participate in very similar pathways. The two diseases are highly comorbid ($RR = 1.8$). Yet, an association of psoriasis and vasculitis has only rarely been reported in the literature [88–90].

Graves' disease & vasculitis. *Graves' disease*, an autoimmune disease affecting the thyroid, and *vasculitis* are associated with 24 and 36 genes, respectively. Both of their gene associations are mostly derived from GWAS, five of which are shared (C6orf15, HCG22, HLA-S, MICA, NCRNA00171). The two diseases show network localization in the same neighborhood ($s_{\mathrm{AB}} = -0.3$), share significant pathways and are strongly comorbid ($RR = 1.9$). While there is no literature support for an association between the two diseases, the shared inflammatory component suggest common molecular mechanisms.

The following eight disease pairs exhibit strong network-overlap, despite the fact that they lack known shared genes:

Lymphoma & myocardial infarction. While *lymphoma*, malignancies that develop from lymphocytes, and *myocardial infarction*, a cardiovascular disease, have no known shared disease genes, they have strongly overlapping modules in the interactome ($s_{\mathrm{AB}} = -0.24$). We therefore expect that these two diseases are related at the molecular level. Indeed, we find that SMARCA4, associated with myocardial infarction, interacts with three lymphoma disease genes (ALK, MYC and NFKB2). Cancer cells frequently depend on chromatin regulatory activities to maintain a malignant phenotype. It has been shown that leukemia cells require the SWI/SNF chromatin remodeling complex containing the SMARCA4 protein as the catalytic subunit for their survival and aberrant self-renewal potential [43]. The molecular relatedness between the two diseases is further supported by a high comorbidity ($RR = 2.1$) and the clinical finding that intravascular large cell lymphoma

affects and obstructs the small vessels of the heart [44].

Glioma & myocardial infarction. A similar molecular relation is also found between *myocardial infarction* and *glioma*, a central nervous system tumor ($s_{\mathrm{AB}} = -0.2$). Here, SMARCA4 interacts with the glioma disease genes APC, PPARG and CTNNB1, offering a hint of the potential molecular mechanism of the relatedness of the two diseases. The two diseases have a particularly strong comorbidity ($RR = 6.3$).

Metabolic bone disease & myocardial infarction. The *myocardial infarction* module also overlaps with the module of *metabolic bone disease* ($s_{\mathrm{AB}} = -0.14$) despite no common disease genes. While there is only a single direct protein interaction between the two diseases (SMARCA4 interacts with VDA), the two diseases are associated with common pathways: The hemostasis pathway, for example, contains six metabolic bone disease genes (GNAS, SLC7A11, THBD and COL1A1) and two myocardial infarction genes (LRP8 and OLR1). Recently, it has been shown that the expression of OLR1 is promoted by exposure of human T lymphocytes to minimally oxidized low density lipoprotein (LDL) [91]. This link between oxidized lipids and T-lymphocytes indicates a new pathway by which T lymphocytes contribute to bone changes. Furthermore, COL1A1 knockout mice were shown to have lower mean arterial and systolic blood pressure, reduced ventricular systolic function, and decreased diastolic function, compared with wild-type littermates [92].

Glioma & gout. The interactome overlap ($s_{\mathrm{AB}} = -0.14$) of the diseases *glioma* and *gout*, the latter a metabolic disease associated with arthritis and kidney stress, is mediated by two protein interactions (the glioma genes BRAF and SGK1 interact with the gour genes GOPC and GRID2, respectively). This association is further supported by a recently reported upregulation of SGK1 in tumor tissue [93] and a high comorbidity between the two diseases ($RR = 2.4$).

Liver cirrhosis & spondylitis. The overlap ($s_{\mathrm{AB}} = -0.2$) between *liver cirrhosis* and *spondylitis*, an inflammatory disease of the axial skeleton, is supported in part by the clinical observation that vertebral fractures occur in 12%-55% of patients with cirrhosis [94]. At the molecular level, the two diseases have three interacting protein pairs (TNFRSF1A and ERAP1; IL12A and IL12B; IL12RB2 and IL12B).

Chronic lymphocytic leukemia & motor neuron disease. While the two diseases do not have common disease genes, there are six direct interactions between the respective disease proteins (ATM and VCP; BCL2 and SOD1; ATM and FUS; SP110 and SMN1; MYC and TIAM1; BCL2

and SMN1). Indeed, the *chronic lymphocytic leukemia* gene BCL2 and its analogs have been found to protect different classes of neurons from apoptosis in several experimental situations [95]. On the other hand, the gene TIAM1, associated with *motor neuron disease*, has been shown to exhibit decreased transcription in unstimulated peripheral chronic lymphocytic leukemia cells [96]. A recent case of a female patient with chronic lymphocytic leukemia presenting with lower motor neuron disease offers clinical evidence for the relation between the two diseases [97].

Albuminuria & bronchial and respiratory diseases. *Albuminuria*, a form of proteinuria characterized by the presence of albumin in the urine and indicative of renal damage, overlaps with bronchial and respiratory diseases, such as *asthma* ($s_{\mathrm{AB}} = -0.13$). At the molecular level, two albuminuria-associated proteins (KCND2 and PARD3B) interact directly with two proteins associated with respiratory diseases (DPP10 and SMAD3). The diseases also exhibit an elevated comorbidity ($RR = 1.23$). Furthermore, it has been suggested that poor lung function, particularly the restrictive pattern, is related to kidney damage as well as atherosclerosis [98].

Proteinuria & myeloproliferative disorders. Other unsuspected diseases found in the network neighborhood of *proteinuria* are *myeloproliferative disorders*, a group of conditions wherein blood cells grow abnormally in the bone marrow ($s_{\mathrm{AB}} = -0.137$). In support of this association, nephrotic-range proteinuria appears to be a late complication of myeloproliferative disorder [99].

10 A disease module approach for the interpretation of GWAS results

In the following we show how our network-based approach can be used to enhance the interpretation of GWAS data by identifying proteins with promising disease associations even though they do not meet the conservative significance criteria of GWAS. GWAS typically identifies a few statistically significant loci, together with hundreds of weaker loci eliminated by the large multiple hypothesis correction. Protein-interaction data, combined with various other -omics or functional association data, have been used previously to prioritize disease gene candidates, resulting in tools like GeneMania [100] and others [8, 9, 17, 19–28, 101, 102]. Here, we tested whether the observed localization of disease proteins in the interactome could help identify the potential biological role of statistically less significant GWAS loci. If we consider a combination of OMIM and only GWAS genes with genome-wide significance (p-value $\leq 5 \times 10^{-8}$), the observable *type 2 diabetes* (T2D)

module has only nine genes, three of which are derived from GWAS. When we add GWAS genes of lower significance in the order of their p-value, we observe that while most GWAS genes are isolated, a few connect to the module (Figure S16a), resulting in distinct jumps in the module size S and its z-score obtained from simulations using randomly chosen genes from the network. These jumps help us single out GWAS candidate genes that may play an important role in the integrity of the T2D disease module (Figure S16a), an approach similar to that used in functional brain networks [103]. For example, at step 7 calmodulin 2 (CALM2), a gene lacking genome-wide significance (7×10^{-7}), links three previously separated OMIM genes to the observable module. CALM2 has been identified in a GWAS analysis of diabetic complications, but its central position in the module suggest an important role in T2D. Indeed, CALM2 is expressed in pancreatic beta cells and is activated by glucose, resulting in exocytosis of insulin [104]. The cluster connected by CALM2 links apolipoprotein B (APOB) to the module, which, once again alone lacks genome-wide significance (p-value $= 8 \times 10^{-8}$), suggesting that its known role in affecting circulating lipid levels also plays a role in genetic predisposition to T2D, a long debated hypothesis [105–107]. Equally interesting is the addition of IDE at step 9. Indeed, variations near HHEX (hematopoietically-expressed homeobox protein)/IDE/KIF11 have shown the strongest association with T2D in a meta-analysis, with HHEX believed to be the disease-related gene. Yet, the incorporation of IDE in the module signifies its potential importance to T2D. Finally, protein kinase C α (PRKCA), which again lacks genome-wide significance, helps link the potassium inward rectifying channel subfamily J (KCNJ11) to the module, a gene of genome-wide significance regulated by PRKCA, which is itself connected to ABCC8, another OMIM gene associated with T2D.

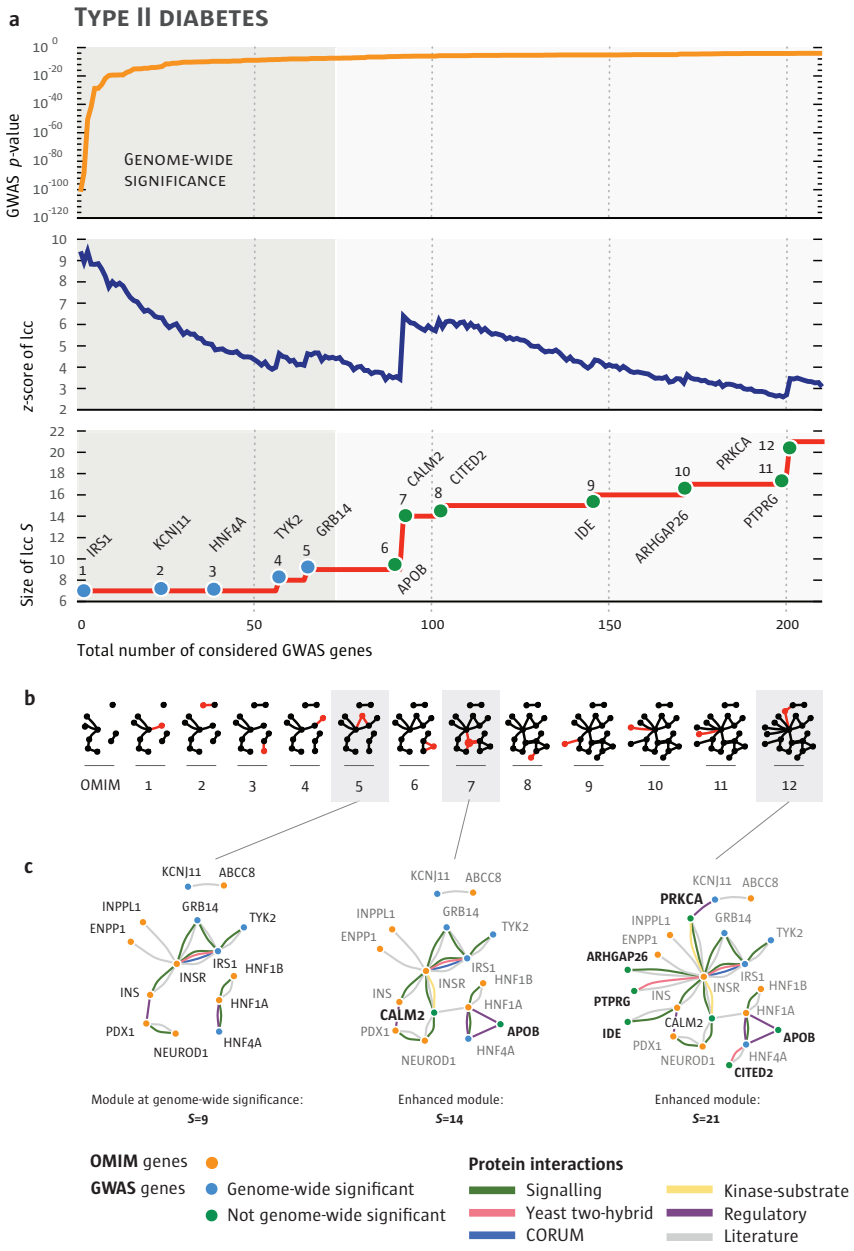


Figure S16: **Identifying biologically relevant GWAS genes. a,** GWAS studies yield a plethora of genes with moderate effect size, lacking genome-wide significance. For type 2 diabetes, starting from an initial module of six connected OMIM genes, we add GWAS genes in the order of their p-value (top), monitoring the growth of the module (bottom) and its statistical significance compared to randomly chosen genes from the network (middle). **b,c,** Five of the 77 genome-wide significant genes connect to the initial OMIM module. With the addition of CALM2 two initially disconnected clusters merge, leading to a distinct jump in the module size and its significance. After the consideration of 200 GWAS genes, another disconnected cluster is integrated, which contains the genome-wide significant gene KCNJ11 and the OMIM gene ABCC8.

11 Supplementary Data

The following datasets are available together with this publication. The full data can also be downloaded from the website `www.barabasilab.com`.

Human interactome. Table with the interactome as described in Section 1.1.

Disease-gene associations. Table with 299 diseases and their associated genes as described in Section 1.2.

Network properties of all diseases. Table with the observable module size S_i, the mean shortest distance $\langle d_s \rangle$, as well as their respective statistical significance for all 299 diseases.

Network properties of all disease pairs. Table with network-based mean distance $\langle d_{\mathrm{AB}} \rangle$, separation s_{AB} and the respective statistical significance according to two randomization schemes (see section 2.2.1) for all 44,551 disease pairs.

Source code. The source code to compute network localization and separation is available as a python package.

References

[51] Matys, V. *et al.* Transfac®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31**, 374–378 (2003).

[52] Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).

[53] Yu, H. *et al.* Next-generation sequencing to generate interactome datasets. *Nature Meth.* **8**, 478–480 (2011).

[54] Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437**, 1173–1178 (2005).

[55] Center for Cancer Systems Biology. Hi-2012 prepublication. `http://interactome.dfci.harvard.edu/H_sapiens/`.

[56] Aranda, B. *et al.* The intact molecular interaction database in 2010. *Nucleic Acids Res.* **38**, D525–D531 (2010).

[57] Ceol, A. *et al.* Mint, the molecular interaction database: 2009 update. *Nucleic Acids Res.* **38**, D532–D539 (2010).

[58] Stark, C. *et al.* The biogrid interaction database: 2011 update. *Nucleic Acids Res.* **39**, D698–D704 (2011).

[59] Prasad, T. K. *et al.* Human protein reference database2009 update. *Nucleic Acids Res.* **37**, D767–D772 (2009).

[60] Lee, D.-S. *et al.* The implications of human metabolic network topology for disease comorbidity. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 9880–9885 (2008).

[61] Ruepp, A. *et al.* Corum: the comprehensive resource of mammalian protein complexes2009. *Nucleic Acids Res.* **38**, D497–D501 (2010).

[62] Hornbeck, P. V. *et al.* Phosphositeplus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* **40**, D261–D270 (2012).

[63] Vinayagam, A. *et al.* A directed protein interaction network for investigating intracellular signal transduction. *Science Signaling* **4**, rs8 (2011).

[64] Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).

[65] Newman, M. E. The structure and function of complex networks. *SIAM review* **45**, 167–256 (2003).

[66] Berriz, G. F., Beaver, J. E., Cenik, C., Tasan, M. & Roth, F. P. Next generation software for functional trend analysis. *Bioinformatics* **25**, 3043–3044 (2009).

[67] Katz, D., Baptista, J., Azen, S. & Pike, M. Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics* 469–474 (1978).

[68] Barraez, D., Boucheron, S., De La Vega, W. F., Fernandez, W. & La, D. The giant component is normal. *Combin Probab Comput* **5133** (1999).

[69] Bender, E. A. & Canfield, E. R. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A* **24**, 296–307 (1978).

[70] Bollobás, B. Random graphs. In *Graph Theory*, 123–145 (Springer, 1979).

[71] Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296**, 910–913 (2002).

[72] Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 5116–5121 (2001).

[73] Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 9440–9445 (2003).

[74] Stumpf, M. P., Wiuf, C. & May, R. M. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 4221–4224 (2005).

[75] Guimerà, R. & Sales-Pardo, M. Missing and spurious interactions and the reconstruction of complex networks. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 22073–22078 (2009).

[76] Stumpf, M. P. & Wiuf, C. Incomplete and noisy network data as a percolation process. *J. R. Soc. Interface* **7**, 1411–1419 (2010).

[77] Annibale, A. & Coolen, A. What you see is not what you get: how sampling affects macroscopic features of biological networks. *Interface focus* **1**, 836–856 (2011).

[78] Callaway, D. S., Newman, M. E., Strogatz, S. H. & Watts, D. J. Network robustness and fragility: Percolation on random graphs. *Phys. Rev. Lett.* **85**, 5468 (2000).

[79] Newman, M. E., Strogatz, S. H. & Watts, D. J. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118 (2001).

[80] Cohen, R., Erez, K., Ben-Avraham, D. & Havlin, S. Resilience of the internet to random breakdowns. *Phys. Rev. Lett.* **85**, 4626 (2000).

[81] Dorogovtsev, S. N. & Mendes, J. F. *Evolution of networks: From biological nets to the Internet and WWW* (Oxford University Press, 2003).

[82] Erdős, P. & Rényi, A. On random graphs. *Publicationes Mathematicae Debrecen* **6**, 290–297 (1959).

[83] Fortunato, S. Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010).

[84] Ahn, Y.-Y., Bagrow, J. P. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466**, 761–764 (2010).

[85] Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).

[86] Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).

[87] Goto, T. *et al.* Primary biliary cirrhosis associated with membranous glomerulonephritis. *Internal Medicine* **38**, 22 (1999).

[88] Wong, S. & Marks, R. Cutaneous vasculitis in psoriasis. *Acta dermato-venereologica* **74**, 57 (1994).

[89] Demitsu, T. *et al.* Pustular vasculitis with clinical feature of pustular psoriasis and sternoclavicular hyperostosis. *Dermatology* **186**, 213–216 (2009).

[90] Moreno, J. *et al.* Psoriasis, vasculitis and methotrexate. *Journal of the European Academy of Dermatology and Venereology* **17**, 466–468 (2003).

[91] Graham, L. S. *et al.* Oxidized lipids enhance rankl production by t lymphocytes: implications for lipid-induced bone loss. *Clin. Immunol.* **133**, 265–275 (2009).

[92] Nong, Z. *et al.* Type I collagen cleavage is essential for effective fibrotic repair after myocardial infarction. *Am. J. Pathol.* **179**, 2189–2198 (2011).

[93] Simon, P. *et al.* Differential regulation of serum-and glucocorticoid-inducible kinase 1 (SGK1) splice variants based on alternative initiation of transcription. *Cell. Physiol. Biochem.* **20**, 715–728 (2007).

[94] Collier, J. Bone disorders in chronic liver disease. *Hepatology* **46**, 1271–1278 (2007).

[95] Sagot, Y. *et al.* Bcl-2 overexpression prevents motoneuron cell body loss but not axonal degeneration in a mouse model of a neurodegenerative disease. *J. Neurosc.* **15**, 7727–7733 (1995).

[96] Hofbauer, S. W. *et al.* Tiam1/Rac1 signals contribute to the proliferation and chemoresistance, but not motility, of chronic lymphocytic leukemia cells. *Blood* **123**, 2181–2188 (2014).

[97] Papageorgiou, A., Tzioufas, A. G. & Voulgarelis, M. B-cell chronic lymphocytic leukaemia presenting as lower motor neuron disease and SS. *Rheumatology* **51**, 1338–1340 (2012).

[98] Yoon, J.-H., Won, J.-U., Ahn, Y.-S. & Roh, J. Poor lung function has inverse relationship with microalbuminuria, an early surrogate marker of kidney damage and atherosclerosis: The 5th korea national health and nutrition examination survey. *PloS one* **9**, e94125 (2014).

[99] Said, S. M. *et al.* Myeloproliferative neoplasms cause glomerulopathy. *Kidney intern.* **80**, 753–759 (2011).

[100] Warde-Farley, D. *et al.* The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **38**, W214–W220 (2010).

[101] Ray, M., Ruan, J., Zhang, W. *et al.* Variations in the transcriptome of alzheimer's disease reveal molecular networks involved in cardiovascular diseases. *Genome Biol* **9**, R148 (2008).

[102] Navlakha, S., Gitter, A. & Bar-Joseph, Z. A network-based approach for predicting missing pathway interactions. *PLoS Comp. Biol.* **8**, e1002640 (2012).

[103] Gallos, L. K., Makse, H. A. & Sigman, M. A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 2825–2830 (2012).

[104] Gloyn, A. *et al.* Human calcium/calmodulin-dependent protein kinase ii gamma gene (camk2g): cloning, genomic structure and detection of variants in subjects with type ii diabetes. *Diabetologia* **45**, 580–583 (2002).

[105] Bernard, S. *et al.* Relation between xba1 apolipoprotein b gene polymorphism and cardiovascular risk in a type 2 diabetic cohort. *Atherosclerosis* **175**, 177–181 (2004).

[106] Ukkola, O., Salonen, J. & Kesäniemi, Y. A. Role of candidate genes in the lipid responses to intensified treatment in type 2 diabetes. *Journal of endocrinological investigation* **28**, 871–875 (2005).

[107] Biddinger, S. B. *et al.* Hepatic insulin resistance is sufficient to produce dyslipidemia and susceptibility to atherosclerosis. *Cell metabolism* **7**, 125–134 (2008).