Network medicine: a network-based approach to human disease

Albert-László Barabási[*‡§], *Natali Gulbahce*[*‡||] *and Joseph Loscalzo*[§]

Abstract | Given the functional interdependencies between the molecular components in a human cell, a disease is rarely a consequence of an abnormality in a single gene, but reflects the perturbations of the complex intracellular and intercellular network that links tissue and organ systems. The emerging tools of network medicine offer a platform to explore systematically not only the molecular complexity of a particular disease, leading to the identification of disease modules and pathways, but also the molecular relationships among apparently distinct (patho)phenotypes. Advances in this direction are essential for identifying new disease genes, for uncovering the biological significance of disease-associated mutations identified by genome-wide association studies and full-genome sequencing, and for identifying drug targets and biomarkers for complex diseases.

*Center for Complex Networks Research and Department of Physics, Northeastern University, 110 Forsyth Street, 111 Dana Research Center, Boston, Massachusetts 02115, USA.
‡Center for Cancer Systems Biology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, USA.
§Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 75 Francis Street, Boston, Massachusetts 02115, USA.
||Department of Cellular and Molecular Pharmacology, University of California, 1700 4th Street, Byers Hall 309, Box 2530, San Francisco, California 94158, USA.
Correspondence to A.-L.B.
e-mail: alb@neu.edu
doi:10.1038/nrg2918

Most cellular components exert their functions through interactions with other cellular components, which can be located either in the same cell or across cells, and even across organs. In humans, the potential complexity of the resulting network — the human interactome — is daunting: with ~25,000 protein-coding genes, ~1,000 metabolites and an undefined number of distinct proteins[1] and functional RNA molecules, the number of cellular components that serve as the nodes of the interactome easily exceeds 100,000. The number of functionally relevant interactions between the components of this network, representing the links of the interactome, is expected to be much larger[2].

This inter- and intracellular interconnectivity implies that the impact of a specific genetic abnormality is not restricted to the activity of the gene product that carries it, but can spread along the links of the network and alter the activity of gene products that otherwise carry no defects. Therefore, an understanding of a gene's network context is essential in determining the phenotypic impact of defects that affect it[3,4]. Following on from this principle, a key hypothesis underlying this Review is that a disease phenotype is rarely a consequence of an abnormality in a single effector gene product, but reflects various pathobiological processes that interact in a complex network. A corollary of this widely held hypothesis is that the interdependencies among a cell's molecular components lead to deep functional, molecular and causal relationships among apparently distinct phenotypes.

Network-based approaches to human disease have multiple potential biological and clinical applications. A better understanding of the effects of cellular interconnectedness on disease progression may lead to the identification of disease genes and disease pathways, which, in turn, may offer better targets for drug development. These advances may also lead to better and more accurate biomarkers to monitor the functional integrity of networks that are perturbed by diseases as well as to better disease classification. Here we present an overview of the organizing principles that govern cellular networks and the implications of these principles for understanding disease. These principles and the tools and methodologies that are derived from them are facilitating the emergence of a body of knowledge that is increasingly referred to as network medicine[5–7].

The human interactome

Although much of our understanding of cellular networks is derived from model organisms, the past decade has seen an exceptional growth in human-specific molecular interaction data[8]. Most attention has been directed towards molecular networks, including protein interaction networks, whose nodes are proteins that are linked to each other by physical (binding) interactions[9,10]; metabolic networks, whose nodes are metabolites that are linked if they participate in the same biochemical reactions[11–13]; regulatory networks, whose directed links represent either regulatory relationships between a transcription factor and a gene[14], or post-translational

Box 1 | **Biological network maps and interaction resources**

Although the bulk of research on biological networks has focused on *Escherichia coli* and *Saccharomyces cerevisiae*, following the Human Genome Project the amount of data pertaining to networks in the human cells exceeds in richness and diversity the data that are available for model organisms. In the following, we briefly discuss the most studied network maps and their limitations, but we remind the reader to exercise caution as we are describing a rapidly changing landscape. The links and references to pertinent databases are available online.

Protein–protein interaction networks

The past 5 years has seen significant efforts towards obtaining comprehensive protein interaction maps. High-throughput yeast two-hybrid maps for humans have been generated by several groups[2,9,10,122], yielding more than 7,000 binary interactions. The immunoprecipitation and high-throughput mass spectrometry technique, which identifies co-complexes, is now being applied to humans as well[123]. There have also been major efforts to curate the interactions that have been validated individually in the literature into databases[124] such as the Münich Information Center for Protein Sequence (MIPS) protein interaction database, the Biomolecular Interaction Network Database (BIND), the Database of Interacting Proteins (DIP), the Molecular Interaction database (MINT), and the protein Interaction database (IntAct). More recent protein–protein interaction curation efforts, including the Biological General Repository for Interaction Datasets (BioGRID) and the Human Protein Reference Database (HPRD), have attempted larger-scale curation of data. Additionally, the STRING database contains known and predicted protein–protein interactions. Despite these extensive curation efforts, the existing maps are considered incomplete[2], and the literature-based data sets, although richer in interactions, are prone to investigative biases[25] as they contain more interactions for the more explored disease proteins[41].

Metabolic networks

The metabolic network maps are probably the most comprehensive of all biological networks. Databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the Biochemical Genetic and Genomics knowledgebase (BIGG) contain the metabolic network of a wide range of species. Recently, Duarte *et al.*[13] published a comprehensive literature-based genome-scale metabolic reconstruction of human metabolism, with 2,766 metabolites and 3,311 metabolic and transport reactions. An independent manual construction by Ma *et al.*[125] contains nearly 3,000 metabolic reactions, organized into about 70 human-specific metabolic pathways.

Regulatory networks

Mapping of the human regulatory network is in its infancy, making this network perhaps the most incomplete among all biological networks. Data generated by experimental techniques, such as chromatin immunoprecipitation (ChIP) followed by microarrays (ChIP–chip) and ChIP followed by sequencing (ChIP–seq), have started to be collected in databases such as the Universal Protein Binding Microarray Resource for Oligonucleotide Binding Evaluation (UniPROBE) and JASPAR. Literature-curated and predicted protein–DNA interactions have been compiled in various databases, such as TRANSFAC and the B-cell interactome (BCI). Human post-translational modifications can be found in databases such as Phospho.ELM, PhosphoSite, phosphorylation site database (PHOSIDA), NetPhorest and the CBS prediction database.

RNA networks

RNA networks can refer to networks containing RNA–RNA or RNA–DNA interactions. Recently, with the increased understanding of microRNAs' role in disease[94], microRNA–gene networks have been constructed using predicted microRNA targets available in databases such as TargetScan, PicTar, microRNA, miRBase and miRDB. The number of experimentally supported targets is also increasing, and they are now compiled in databases such as TarBase and miRecords.

Node (or vertex)
A system component that, by interacting with other components, forms a network. In biological networks, nodes can denote proteins, genes, metabolites, RNA molecules or even diseases and phenotypes.

modifications, such as those between a kinase and its substrates[15]; and RNA networks, which capture the role of interactions between regulatory RNAs, such as small non-coding microRNAs (miRNAs)[16] and small interfering RNAs (siRNAs)[17], and DNA in regulating gene expression. In parallel, an increasing number of studies rely on phenotypic networks, including co-expression networks, in which genes with similar co-expression patterns are linked[18]; and genetic networks, in which two genes are linked if the phenotype of a double mutant differs from the expected phenotype of two single mutants[19,20]. Typically, the links of a phenotypic network reflect some of the pathways in the underlying molecular networks. Finally, both the nodes and the interactions discussed above need to be evaluated in the context of tissue specificity[21–23], as they may exert a functional role in the context of only selected tissues.

The first step in exploring the interplay between networks and human diseases is to assess how comprehensive and accurate the current molecular and phenotypic network maps are for humans (BOX 1). The past few years have witnessed systematic efforts to increase the coverage of human interactome maps, to estimate the interactome size and to correct for known biases[2,24,25]. Still, human interactome maps remain incomplete and noisy, a fact that needs to be taken into account when using them to study diseases. Furthermore, much of the current work, and hence this Review, focuses on intracellular networks, ignoring, owing to lack of systematic data, the molecular networks that connect cells, tissues and organ systems[4,26].

Properties of disease networks

Network medicine relies on a series of advances in network theory[27–32] that have provided insights into the properties of biological networks more generally. These studies have indicated that networks operating in biological, technological or social systems are not random, but are characterized by a core set of organizing principles (BOX 2). Understanding diseases in the context of these network principles allows us to address some fundamental properties of the genes that are involved in disease. Indeed, only about 10% of human genes have a known disease association[33] (FIG. 1a); thus, do disease genes have unique, quantifiable characteristics that distinguish them from other genes? From a network perspective, this question translates as follows: are disease genes placed randomly on the interactome, or are there detectable correlations between their location and their network topology? The search for answers has led to a series of hypotheses that tie the interactome to human diseases. These are summarized in BOX 3, and in the remainder of this article we will discuss the validity and applications of those hypotheses that have received the most attention.

Location of disease genes within networks. An unexpected property of biological networks is the emergence of a few highly connected nodes, often called hubs (BOX 2), suggesting that the proteins represented by these hubs must have a special biological role. Indeed, evidence from model organisms indicates that hub proteins tend to be encoded by essential genes[34], and that genes encoding hubs are older and evolve more slowly than genes encoding non-hub proteins[35–37]. The deletion of genes encoding hubs also leads to a larger number of phenotypic outcomes than for other genes[25]. Although the strength of evidence for some of these effects is still debated[25,38], by virtue of the many interactions it has, the absence of a hub would be expected to affect many more other proteins than would the absence of a non-hub protein.

Link (or edge)
A link represents the interactions between the nodes of a network. In biological systems, interactions can correspond to protein–protein binding interactions or metabolic coupling, or they may represent connections between diseases based on a common genetic origin or shared phenotypic characteristics.

This assumption has led to the hypothesis that, in humans, hubs should typically be associated with disease genes. Some studies support this hypothesis, such as the finding that the protein products of genes that are upregulated in squamous cell carcinoma of the lung tend to have a higher degree than do proteins with levels that are unaffected[39]. In a separate study, 346 proteins that are implicated in cancer were found to have, on average, twice as many interaction partners as did non-cancer proteins[40]. Moving beyond cancer, one study[41] found that disease proteins in the OMIM Morbid Map[33] have more protein–protein interactions than do non-disease proteins.

Box 2 | **Elements of network theory**

An important realization of the past decade is that networks in natural, technological and social systems are not random, but follow a series of basic organizing principles in their structure and evolution that distinguish them from randomly linked networks. In the following, we summarize the aspects of network theory that pertain to biological networks. For a more detailed exposition, see REFS 27–32. Although these principles were found to apply to a range of networks, in the context of this Review, they refer to biological networks, seen as nodes (for example, proteins, metabolites or diseases) connected by links (for example, protein–protein interactions, metabolic reactions or shared genes), as discussed throughout.

Modules
Most networks show a high degree of clustering, implying the existence of topological modules that represent highly interlinked local regions in the network. Although the identification of such modules can be computationally challenging, a wide array of network-clustering tools have emerged over the past few years[47–50].

Degree distribution and hubs
In a random network, most nodes have approximately the same number of links, and highly connected nodes (hubs) are rare. The fraction of links with a given degree, called the degree distribution, follows the well-known Poisson distribution. By contrast, many real networks, including human protein–protein interaction and metabolic networks, are scale free[126], which means that the degree distribution has a power-law tail; that is, the degree distribution $P(k)$, with degree k, follows $P(k) \sim k^{-\gamma}$, where γ is called the degree exponent. The most noticeable consequence of this property is the presence of a few highly connected hubs that hold the whole network together[34]. Researchers often refer to the 20% of nodes in a network with the highest degree as 'hubs', but this definition is arbitrary, as the scale-free property implies that these networks do not have an intrinsic 'scale' — that is, an inherent threshold beyond which nodes are hubs. The biological role and dynamical behaviour of hubs allowed their classification into 'party' hubs, which function inside modules and coordinate specific cellular processes, and 'date' hubs, which link together rather different processes and organize the interactome[66,127].

Small-world phenomena
Most complex networks (including random networks) have the small-world property, which means that there are relatively short paths between any pair of nodes[128]. This observation means that most proteins (or metabolites) are only a few interactions (or reactions) from any other proteins (metabolites)[11,12,34]. Therefore, perturbing the state of a given node can affect the activity of most nodes in their vicinity as well as of the behaviour of the network itself.

Motifs
Some subgraphs (a group of nodes that link to each other, forming a small subnetwork within a network) in biological networks appear more (or less) frequently than expected given the network's degree distribution. Such subgraphs are often called motifs[129], and they are likely to be associated with some optimized biological function (for example, negative feedback loops, positive feedforward loops, bifans or oscillators).

Betweenness centrality
Nodes with a high betweenness centrality (a measure of the number of shortest paths that go through each node) are often called bottlenecks. In networks with directed edges, such as regulatory networks, bottlenecks tend to correlate with essentiality[130].

However, not all essential genes are disease genes in humans. Mutations in genes that are essential in early development cannot propagate in the population, as functional changes in them often lead to first-trimester spontaneous abortions (embryonic lethality); by contrast, individuals can tolerate disease-causing mutations for much longer, often past their reproductive age. This suggests that most disease genes in humans will in fact not be essential genes[42,43]. So, what is the relationship among essential genes, hubs and genes that cause disease in humans? Goh *et al.*[42] found that essential genes that are not associated with disease show a strong tendency to be associated with hubs and are expressed in multiple tissues — that is, they tend to be located at the functional centre of the interactome (FIG. 1). However, non-essential disease genes do not show a tendency to encode hubs, and they tend to be tissue specific and located at the functional periphery of the interactome (FIG. 1b). In summary, in human cells it is the essential genes, and not the disease genes, that encode hubs.

Local clustering of disease genes: disease modules. If a gene or molecule is involved in a specific biochemical process or disease, its direct interactors might also be suspected to have some role in the same biochemical process[44]. In line with this 'local' hypothesis (BOX 3), proteins that are involved in the same disease show a high propensity to interact with each other[42,45]. For example, one group observed 290 physical interactions between the products of genes associated with the same disorder, representing a tenfold increase relative to random expectation[42]. Two other studies found that genes that are linked to diseases with similar phenotypes have a significantly increased tendency to interact directly with each other[41,46]. These observations indicate that, if a few disease components are identified, other disease-related components are likely to be found in their network-based vicinity. That is, we expect that each disease can be linked to a well-defined neighbourhood of the interactome, often referred to as a 'disease module'.

As we try to understand the network-based position of disease genes, we need to distinguish among three distinct phenomena (FIG. 2). A 'topological module' represents a locally dense neighbourhood in a network, such that nodes have a higher tendency to link to nodes within the same local neighbourhood than to nodes outside it. Such modules can be identified using network clustering algorithms that are blind to the function of individual nodes[47–51]. By contrast, a 'functional module' represents the aggregation of nodes of similar or related function in the same network neighbourhood, where function captures the role of a gene in defining detectable phenotypes. Finally, a 'disease module' represents a group of network components that together contribute to a cellular function and disruption of which results in a particular disease phenotype.

In the biological literature, there is a tacit assumption that these three concepts are interrelated: cellular components that form a topological module have closely related functions, thus they also correspond to a functional module, and a disease is a result of the breakdown

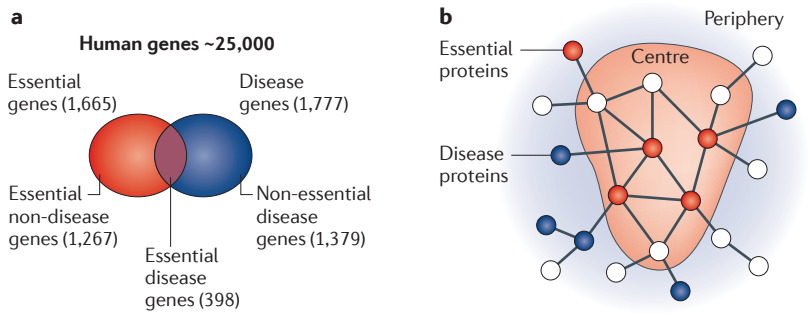


Figure 1 | **Disease and essential genes in the interactome. a** | Of the approximately 25,000 human genes, 2,418 are associated with specific diseases. The figure shows the overlap between the 1,777 disease-associated genes that were known[42] in 2007 and the 1,665 genes that are *in utero* essential, that is, their absence is associated with embryonic lethality. **b** | Schematic diagram of the differences between essential and non-essential disease genes. Non-essential disease genes (shown as blue nodes) are found to segregate at the network periphery, whereas *in utero* essential genes (shown as red nodes) tend to be at the functional centre (encoding hubs and expressed in many tissues) of the interactome. Part **a** is reproduced, with permission, from REF. 42 © (2007) National Academy of Sciences.

of a particular functional module[51], intimating that a functional module is also a disease module. However, several unique characteristics of disease modules are important to bear in mind. First, a disease module may not be identical to, but is likely to overlap with, the topological and/or functional modules. Second, a disease module is defined in relation to a particular disease and, accordingly, each disease has its own unique module. Last, a gene, protein or metabolite can be implicated in several disease modules, which means that different disease modules can overlap. These characteristics aid the disease module identification process, an important step of network medicine (FIG. 3).

The emergence of a disease is therefore viewed as a combinatorial problem in which many different defects and perturbations result in a similar disease phenotype, provided that they alter the activity of the disease module. Such combinatorial disease mechanisms are well documented in cancer[52], but the utility of the disease module hypothesis extends beyond polygenic diseases and is important even in some monogenic diseases. For example, sickle cell disease, a classic Mendelian disorder, is caused by a single point mutation at position 6 of the β-chain of haemoglobin. Still, this simple biochemical phenotype and its corresponding monogenotype do not yield a single pathophenotype: individuals with sickle cell disease can present with painful crises, osteonecrosis, acute chest syndrome, stroke, profound anaemia or mild asymptomatic anaemia. Thus, the underlying disease module is likely to include all disease-modifying genes (for example, haemoglobin F) that mediate various epigenetic, transcriptional and post-translational phenomena. An important step of network-based approaches to disease is, therefore, to identify the disease module for the pathophenotype of interest, which, in turn, can guide further experimental work towards uncovering the disease mechanism, predicting disease genes and influencing drug development.

Degree
The degree of a node is the number of links that connect to it. The degree of a protein could represent the number of proteins with which it interacts with, whereas the degree of a disease may represent the number of other diseases that are associated with the same gene or that have a common phenotype.

Module (or community)
A dense subgraph on the network that often represents a set of nodes that have a joint role. In biology, a module could correspond to a group of molecules that interact with each other to achieve some common function.

Predicting disease genes

Disease-associated genes have generally been identified using linkage mapping or, more recently, genome-wide association (GWA) studies[53]. Both methodologies can suggest large numbers of disease-gene candidates, but identifying the particular gene and the causal mutation remains difficult. Recently, a series of increasingly sophisticated network-based tools have been developed to predict potential disease genes; these tools can be loosely grouped into three categories, as discussed below (FIG. 4).

Linkage methods. These methods assume that the direct interaction partners of a disease protein are likely to be associated with the same disease phenotype[45,54–56]. Indeed, for one disease locus, the set of genes within the locus whose products interacted with a known disease protein were shown to be tenfold enriched in true disease-causing genes[45]. By also considering cellular localization, this approach led to a 1,000-fold enrichment over a random selection. On this basis, the authors predicted and confirmed the involvement of Janus kinase 3 (JAK3) in severe combined immunodeficiency syndrome owing to its interaction with known disease-associated proteins.

Disease module-based methods. A second set of methods assumes that all cellular components that belong to the same topological, functional or disease module have a high likelihood of being involved in the same disease[57,58]. These methods start with identifying the disease modules and inspecting their members as potential disease genes. Disease modules can be identified on the basis of currently available data using bioinformatics approaches (FIG. 3). Briefly, this strategy involves constructing the interactome in the tissue and cell line of interest and identifying a subnetwork, or disease module, that contains most of the disease-associated genes. Disease modules are then validated by, for example, showing that the genes in a module have related functions or have correlated expression patterns.

Variants of this methodology have been applied to a wide range of diseases and pathophenotypes, including several different types of cancer[59–66], neurological diseases[67–69], cardiovascular diseases[68,70], systemic inflammation[71,72], obesity[73–75], asthma[76], type 2 diabetes[77] and chronic fatigue syndrome[78]. For example, Taylor *et al.*[66] identified disease-associated protein interaction modules for adenocarcinoma of the breast, providing useful indicators for predicting breast cancer outcome. Similarly, Chen *et al.*[73] identified subnetworks in liver and adipose tissues that contain genes for which variants associated with obesity and diabetes have been identified. The results confirmed a previously proposed connection between obesity and a macrophage-enriched metabolic subnetwork, validating three previously unknown genes, lipoprotein lipase (*Lpl*), β-lactamase (*Lactb*) and protein phosphatase, Mg^{2+}/Mn^{2+} dependent, 1L (*Ppm1l*), as obesity genes in transgenic mice. The disease module-based approach has also been useful in exploring pathogen-induced phenotypes[79–81] (N.G. *et al.*, unpublished data).

Comorbidity
Comorbidity implies the presence of one or more disorders (or diseases) in addition to a primary disease or disorder that the patient has. Comorbidity may hide causal effects, when one disease enhances the emergence of some other disease, such as the much-studied comorbidity between diabetes and obesity.

An area that is ripe for network-based approaches is the bacterial microbiome (and other metagenomes) and its relationship to human disease[82].

Often the rate-limiting step in mapping a disease module is the low coverage of the available cellular interaction maps in the vicinity of the known disease components, which means that additional experimental efforts are needed to identify relevant interactions. This approach was successfully applied to several diseases, including Huntington's disease[83], spinocerebellar ataxia[84], breast cancer[85] and schizophrenia[86]. For example, starting from 23 known ataxia-causing genes, Lim *et al.*[84] used yeast two-hybrid assays to map their interactions with other human proteins; the interactions of this second group of proteins were then used to build a dense subnetwork that was two degrees removed from the known ataxia genes. A member of the predicted ataxia disease module, puratrophin 1, a common binding partner to many of the known ataxia genes, which were not previously recognized as having any commonality, was later shown to lead to ataxia-like phenotypes in mice following its deletion[87].

Diffusion-based methods. A final group of methods aims to identify the pathways that are closest to the known disease genes. In these algorithms, 'random walkers' are 'released' from the protein products of the known disease genes, and they are then allowed to diffuse along the links of the interactome, moving to any neighbouring node with equal probability. In this way, one can identify the nodes and links that are closest to the known disease genes, as they will be those that are most often visited by the random walkers. Proteins that interact with several disease proteins will gain a high probabilistic weight, as will those that may not directly interact with any disease proteins but are in close network proximity to them. This approach helps to prioritize proteins and interactions on the basis of their potential involvement in the particular disease. Variants of this methodology have been applied to detecting disease genes related to a wide range of diseases, from diabetes mellitus to prostate cancer and Alzheimer's disease[88,89].

Each of these three methodologies exploits, to an increasing degree, the topological and functional information that is encoded by the interactome. The linkage method involves only pairwise linkage information (local hypothesis; BOX 3), whereas the disease module-based method exploits the full network neighbourhood of disease genes (disease module hypothesis, BOX 3). Finally, diffusion-based methods use the information that is encoded in the full network topology as well as the placement of the known disease genes, thereby simultaneously exploiting both topological and functional modularity (together with the parsimony principle; BOX 3). It is not surprising, therefore, that a recent comparative study found that, on the same data set, linkage-based methods have the least predictive power and diffusion-based methods offer the best predictive performance[57].

In summary, the evidence of the nonrandom placement of disease genes in the interactome has opened a series of opportunities for disease gene predictions. The value of these tools is expected to increase with the wealth of disease gene candidates provided by GWA studies and full-genome sequencing. Indeed, these tools help us to narrow the vast search space offered by the interactome, thus aiding our search for disease mechanisms and, eventually, the development of rational, individualized therapies and potential cures.

Box 3 | Hypotheses of network medicine

Network medicine is based on a series of widely used (and often unspoken) hypotheses and organizing principles that link network structure to biological function and disease. Next, we summarize some of the most frequently used hypotheses, and they are discussed in more detail in the main text.

Hubs
Non-essential disease genes (representing most known disease genes) tend to avoid hubs and segregate at the functional periphery of the interactome. *In utero* essential genes tend to associated with hubs.

Local hypothesis
Proteins involved in the same disease have an increased tendency to interact with each other.

Corollary of the local hypothesis
Mutations in interacting proteins often lead to similar disease phenotypes.

Disease module hypothesis
Cellular components associated with a specific disease phenotype show a tendency to cluster in the same network neighbourhood.

Network parsimony principle
Causal molecular pathways often coincide with the shortest molecular paths between known disease-associated components.

Shared components hypothesis
Diseases that share disease-associated cellular components (genes, proteins, metabolites or microRNAs) show phenotypic similarity and comorbidity.

Human diseasome

The highly interconnected nature of the interactome means that, at the molecular level, it is difficult to consider diseases as being consistently independent of one another. Indeed, different disease modules can overlap, so that perturbations caused by one disease can affect other disease modules. The systematic mapping of such network-based dependencies between pathophenotypes and their disease modules has culminated in the concept of the diseasome[42], which represents disease maps whose nodes are diseases and whose links represent various molecular relationships between the disease-associated cellular components. Uncovering such links between diseases not only helps us understand how different phenotypes, often addressed by different medical subdisciplines, are linked at the molecular level, but can also help us to comprehend why certain groups of diseases arise together. The comorbidity of conditions that are culled from the diseasome offers insights that may yield new approaches to disease prevention, diagnosis and treatment. Diseasome-based approaches could also aid drug discovery, in particular when it comes to the use of approved drugs to treat molecularly linked diseases. Here, we review the construction of such disease maps and the implications of the observed disease associations.

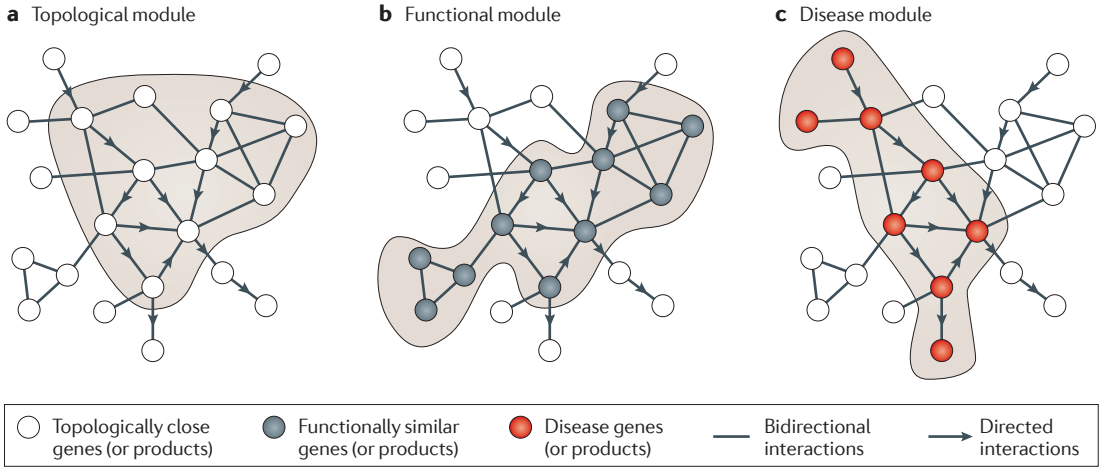


Figure 2 | **Disease modules.** Schematic diagram of the three modularity concepts that are discussed in this Review. **a** | Topological modules correspond to locally dense neighbourhoods of the interactome, such that the nodes of the module show a higher tendency to interact with each other than with nodes outside the module. As such, topological modules represent a pure network property. **b** | Functional modules correspond to network neighbourhoods in which there is a statistically significant segregation of nodes of related function. Thus, a functional module requires us to define some nodal characteristics (shown as grey nodes) and relies on the hypothesis that nodes that are involved in closely related cellular functions tend to interact with each other and are therefore located in the same network neighbourhood. **c** | A disease module represents a group of nodes whose perturbation (mutations, deletions, copy number variations or expression changes) can be linked to a particular disease phenotype, shown as red nodes. The tacit assumption in network medicine is that the topological, functional and disease modules overlap, so that functional modules correspond to topological modules and a disease can be viewed as the breakdown of a functional module.

Shared gene hypothesis and the human disease network. The linkage of a gene to different disease pathophenotypes often indicates that these diseases have a common genetic origin. Motivated by this hypothesis, Goh *et al.*[42] used the gene–disease associations that are collected in the OMIM database to build a network of diseases that are linked if they share one or more genes. In the obtained human disease network (HDN), 867 of 1,284 diseases with an associated gene are connected to at least one other disease, and 516 of them belong to a single disease cluster (FIG. 5). The clustering of nodes of similar colour in FIG. 5, denoting the disease class, reflects the fact that similar pathophenotypes have a higher likelihood of sharing genes than do pathophenotypes that belong to different disease classes. For example, cancers form a tightly interconnected and easily detectable cluster, which is held together by a small group of genes that are associated with multiple cancers.

To determine whether the sharing of genes has consequences for disease occurrence in populations, the comorbidity between linked disease pairs has been examined[90] (FIG. 5). This analysis indicates that a patient is twice as likely to develop a particular disease if that disease shares a gene with the patient's primary disease. But many disease pairs that share genes do not show significant comorbidity. One explanation is that different mutations in the same gene can have different effects on the gene product, and therefore different pathological consequences[91] that are organ and context dependent. Such 'edgetic' alleles affect a specific subset of links in the interactome[92]. Consistent with this view, disease pairs that are associated with mutations that affect the same functional domain of a protein show higher comorbidity than do disease pairs with mutations that occur in different functional domains[90] (FIG. 5).

Shared metabolic pathway hypothesis and the metabolic disease network. An enzymatic defect that affects the flux of one reaction can potentially affect the fluxes of all downstream reactions in the same pathway, leading to disease phenotypes that are normally associated with these downstream reactions. Thus, for metabolic diseases, links that are induced by shared metabolic pathways are expected to be more relevant than are links based on shared genes. In support of this hypothesis, Lee *et al.*[93] constructed a metabolic disease network (MDN) in which two disorders are connected if the enzymes associated with them catalyse adjacent reactions (FIG. 5b). The visually apparent clustering of the MDN mirrors distinct metabolic pathways. For example, purine metabolism consists of 62 reactions associated with 33 diseases, including nucleoside phosphorylase deficiency and congenital dyserythropoietic anaemia, which form a visually distinct cluster. Comorbidity analysis confirms the functional relevance of metabolic coupling: disease pairs that are linked in the MDN have a 1.8-fold increased comorbidity compared to disease pairs that are not linked metabolically[93]. Comorbidity is even more pronounced if the fluxes of the reactions that are catalysed by the respective disease genes are themselves coupled; that is, changes in one flux induce significant changes in the other flux, even if the corresponding reactions are not adjacent.

Edgetic
Edgetic perturbations denote mutations that do not result in the complete loss of a gene product, but affect one or several interactions (and thus functions) of a protein. From a network perspective, an edgetic perturbation removes one or several links, but leaves the other links and the node unaffected.

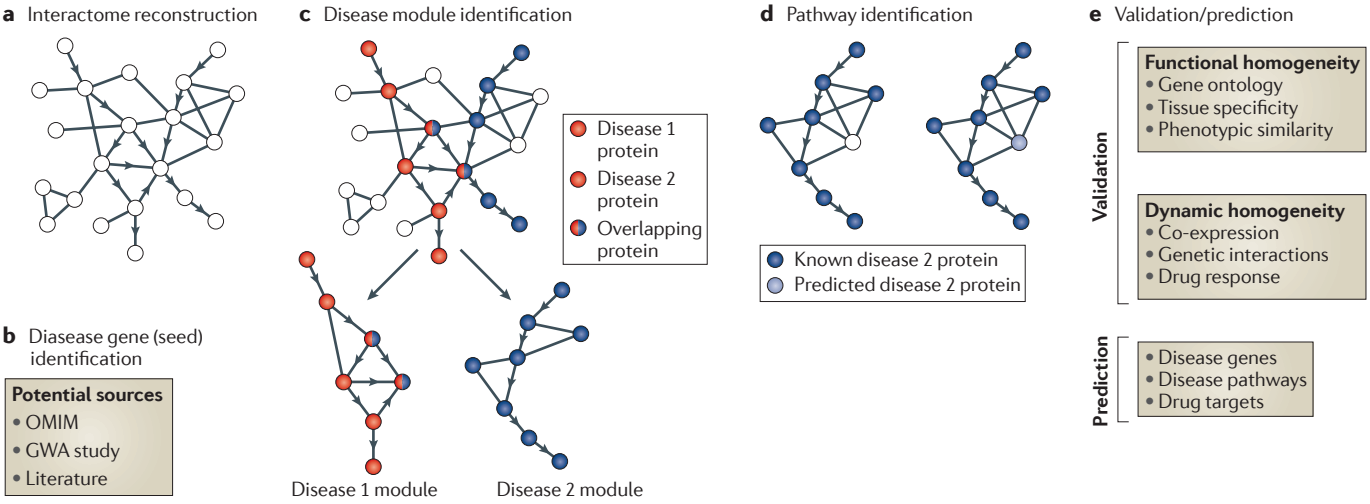


Figure 3 | **Identifying and validating disease modules.** For any specific disease, the identification and validation of disease modules consists of several steps. **a** | Interactome reconstruction merges the most up-to-date information on protein–protein interactions, co-complex memberships, regulatory interactions and metabolic network maps (BOX 1) in the tissue and cell line of interest. These networks are occasionally augmented with phenotypic links, such as co-expression-based relationships[68,73]. We feel that such phenotypic measures are best used later, to test the functional homogeneity of the predicted disease module. **b** | Disease gene (seed) identification collects the known disease-associated genes obtained from linkage analysis, genome-wide association (GWA) studies or other sources, which serve as the seed of the disease module. **c** | In disease module identification the seed genes are placed on the interactome, with the aim of identifying a subnetwork that contains most of the disease-associated components, exploiting both the functional and topological modularity of the network. If such statistically significant agglomeration is detected, then one can use a combination of clustering tools[47–50] to identify the functionally and topologically compact subgraph that contains the most disease components, which thus represents the potential disease module. The closer the phenotypic manifestations of the two diseases (in terms of, for example, organ system, symptoms or drug response), the more significant is the expected overlap between the modules that are associated with two diseases. **d** | Pathway identification can be used in instances in which the number of components contained in the ascertained disease module is so large that it cannot serve as a tractable starting point for further experimental work. In these cases it may be necessary to identify the specific molecular pathways whose disruption may be responsible for the disease phenotype. One typically uses the network parsimony principle (BOX 3) to select the most likely disease pathways, assuming that causal pathways are the shortest paths connecting the known disease components. **e** | During validation disease modules are tested for their functional and dynamic homogeneity. The nature of the validation depends on the tools and data that are available to the investigator; gene expression data can validate the dynamical integrity of the disease module, and GWA studies can be used to test the potential links between the SNPs of the predicted cellular components and the disease phenotype. Finally, the predicted disease genes and pathways (which also serve as potential drug targets) are tested using the available molecular biology tools and animal models.

Shared microRNA hypothesis. Prompted by the increasing evidence of the role of miRNAs in human disease, Lu *et al.*[94] connected disease pairs with associated genes that are targeted by at least one common miRNA molecule. The obtained miRNA-based disease network shows a disease-class-based segregation; for example, cancers share similar associations at the miRNA level, leading to a distinct cancer cluster that differs from, for example, the cluster associated with cardiovascular diseases.

Phenotypic disease networks. One can also link disease pairs on the basis of the directly observed comorbidity between them, thereby obtaining a phenotypic disease network (PDN). For example, Rzhetsky *et al.*[95] inferred the comorbidity links between 161 disorders from the disease history of 1.5 million patients at the Columbia University Medical Center, New York, USA, and Hidalgo *et al.*[96] built a network involving 657 diseases from the disease history of more than 30 million Medicare patients. In these maps, two diseases are connected if their comorbidity exceeds a predefined threshold. The PDN is blind to the mechanism that underlies the observed comorbidity, which may be rooted in molecular-level dependencies (as seen for HDNs, MDNs or miRNA-based disease networks) or in environmental or treatment-related perturbations of the network. Still, PDNs capture disease progression, as patients tend to develop diseases in the network vicinity of diseases that they have already had[96]. Furthermore, patients who are diagnosed with diseases with more links in the PDN show a higher mortality than do those who are diagnosed with diseases that are less well connected[96]. Another use of phenotypic information was suggested by Van Driel *et al.*[97], who used text mining to assign to more than 5,000 human phenotypes in the OMIM database a string of phenotypic features from the medical

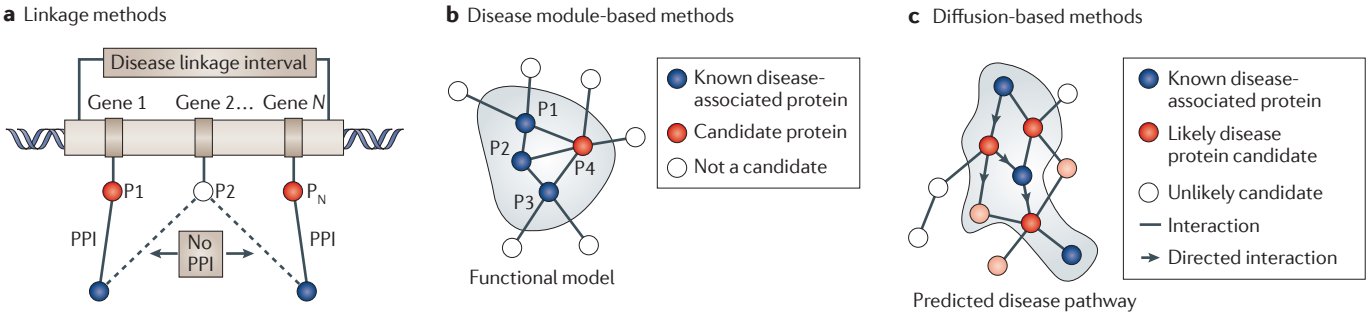


Figure 4 | **Identifying disease gene candidates. a** | Linkage methods. Genes located in the linkage interval of a disease whose protein products (labelled P1, P2, and so on) interact with a known disease-associated protein are considered likely candidate disease genes[45,57]. **b** | Disease module-based methods. Clustering or graph partitioning helps us to uncover functional and potential disease modules in the interactome. The members of such modules are considered candidate disease genes[57,131]. **c** | Diffusion-based methods. Starting from proteins that are known to be associated with a disease, a random walker visits each node in the interactome with a certain probability[88,89]. The outcome of this algorithm is a disease-association score that is assigned to each protein, that is, the likelihood that a particular protein is associated with the disease. PPI, protein–protein interaction.

subject heading vocabulary. The overlap of their phenotypic descriptions was used to link various diseases, and the authors found that phenotypic similarity correlates positively with the molecular signatures of two linked diseases, from relatedness at the level of protein sequence to protein motifs and direct protein–protein interactions between the disease-associated proteins.

Most efforts have focused on the role of a single molecular or phenotypic measure to capture disease–disease relationships (such as shared genes or metabolites), but a comprehensive understanding requires us to inspect multiple sources of evidence, from shared genes to protein–protein interaction-based relationships, shared environmental factors, common treatments, affected tissues and organs[22,23], and phenotypic manifestations. In line with such integrated approaches, Suthram *et al.*[98] built a disease network by linking two diseases for which the same modules were activated in the specific disease states, and Liu *et al.*[99] linked diseases with common environmental influences. Although efforts to understand all causal links between diseases are still in their infancy, they will be essential for a deeper understanding of human disease.

Applying network-based knowledge of disease

Network pharmacology. The first step of rational drug design is to understand the cellular dysfunction that is caused by a disease. By definition, this dysfunction is limited to the disease module, which means that one can reduce the search for therapeutic agents to those that induce detectable changes in module activity. Another application of network pharmacology addresses the fact that, owing to the often unknown interactions between drug targets and other cellular components, drugs whose efficacy was predicted by specific target-binding experiments may not have the same effect *in vivo*. That is, a drug might have more than one binding partner such that its efficacy is determined by its multiple interactions, leading to unwanted side effects[100–102]. Although network-based approaches represent a relatively recent trend in drug discovery, that fact that drug development is clearly affected by intricate network effects[103] suggests that network pharmacology[104] will become an essential component of drug-development strategies. Indeed, the network concepts discussed above have already found their way into drug discovery studies, from attempts to target the hubs with drugs to using modules to identify potential drug targets[105–107].

The promise of network-based approaches in drug discovery is best illustrated in the area of bacterial and human metabolism. Given the relative accuracy of metabolic maps (BOX 1) and the ability to predict flux changes that are induced by drug-altered enzymatic activity in bacteria using flux balance analysis[108] and other flux-based methods[109], the metabolic impact of a hypothetical enzyme-blocking drug can be explored *in silico*. This capability has recently led to the identification and testing of potential new antibacterial agents and the complex system-based responses that they produce[110]. Furthermore, the coupled nature of metabolic fluxes presents the possibility of rescuing a lost metabolic function through inhibiting additional enzymes, which would be selected in order to re-route metabolic activity to compensate for the original loss of function — an intriguing alternative to gene therapy[111].

Single-target drugs may, perhaps, correct some dysfunctional aspects of the disease module, but they could also alter the activity of molecules that are situated in the neighbourhood of the disease module, leading to detectable side effects[100–102]. This network-based view of drug action implies that most disease phenotypes are difficult to reverse through the use of a single 'magic bullet', that is, an intervention that affects a single node in the network[112]. Increasing attention is therefore being given to therapies that involve multiple targets, which may be more effective in reversing the disease phenotype than are single drugs[113]. The efficacy of this approach has been demonstrated by combinatorial therapies for AIDS, cancer and depression, raising an important question: can one systematically identify multiple drug targets that

Figure 5 | **Disease networks. A** | An example of a human disease network (HDN), in which nodes represent diseases. The large panel shows the giant cluster of the obtained disease network (**Aa**). Small clusters of isolated diseases are not shown[42]. Two diseases are linked if they share one or several disease-associated genes, as shown in part **Ab**, involving breast cancer and bone and cartilage cancer[90]. The node colours reflect the class of the diseases that correspond to that node. Cancers appear as blue nodes and neurological diseases appear as red nodes. The node sizes correlate with the number of genes that are known to be associated with the corresponding disease (after REF. 42). Part **Ac** shows the comorbidity between diseases linked in the HDN as measured by the logarithm of relative risk, indicating that, if the disease-causing mutations affect the same module of the shared disease protein, then the comorbidity is higher[90]. **B** | A metabolic disease network (part **Ba**, with an example shown part **Bb**), which links two diseases if they are both associated with enzymes and if these enzymes catalyse reactions that share a metabolite (after REF. 93). Part **Bc** shows that comorbidity between metabolically linked diseases is higher than between those that are not connected, and diseases whose enzymes catalyse reactions that are coupled with each other at the flux level show even higher comorbidity. *AR*, androgen receptor; *ATM*, ataxia telangiectasia mutated; *BRCA*, breast cancer associated; *BARD1*, BRCA1 associated RING domain 1; BIGG, Biochemical Genetic and Genomics knowledgebase; BPGM, 2,3-bisphosphoglycerate mutase; *CHEK2*, CHK2 checkpoint homologue; *CSMF*, chondrosarcoma, extraskeletal myxoid, fused to EWS (also known as *NR4A3*); ENO3, enolase 3 (beta, muscle); *EWSR1*, Ewing sarcoma breakpoint region 1; *ESR1*, oestrogen receptor 1; *EXT1*, exostosin 1; HARP, hypoprebetalipoproteinaemia, acanthocytosis, retinitis pigmentosa and pallidal degeneration; HELLP, haemolytic anaemia, elevated liver enzymes and low platelet count; KEGG, Kyoto Encyclopedia of Genes and Genomes; PGAM2, phosphoglycerate mutase 2 (muscle); *PHB*, prohibitin; *RB1*, retinoblastoma 1; *TAF15*, TATA box-binding protein (TBP)-associated factor, RNA polymerase II; *TSG101*, tumour-susceptibility gene 101; *TP53*, tumour protein 53. Part **Aa** is reproduced, with permission, from REF. 42 © (2007) National Academy of Sciences. Part **Ab** is reproduced, with permission, from REF. 90 © (2009) Macmillan Publishers Ltd. All rights reserved. Part **Bb** is reproduced, with permission, from REF. 93 © (2008) National Academy of Sciences.

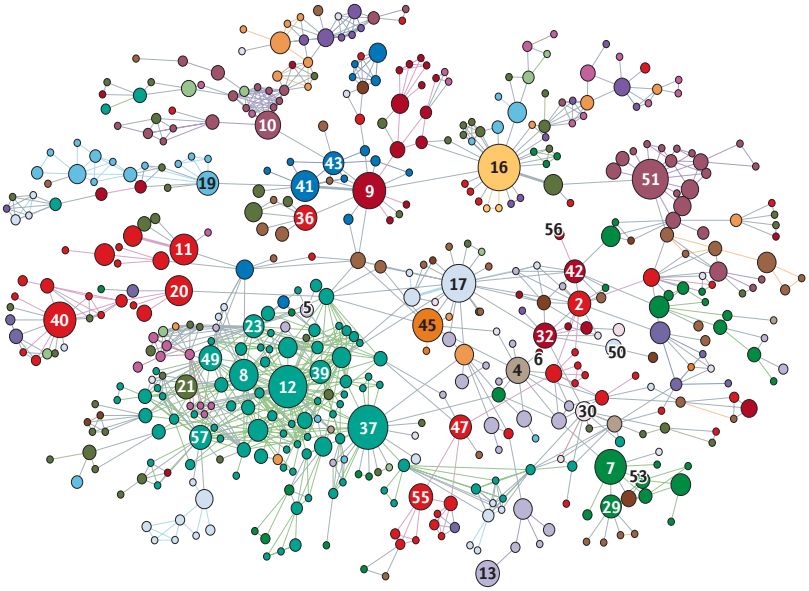
have an optimal impact on the disease phenotype? This is an archetypical network problem, and it has led to the development of methods to identify optimal drug combinations, starting either from the metabolic network[114,115] or from the bipartite network that links compounds to their drug-response phenotypes[116]. Such research has led to potentially safer multi-target combinations for inflammatory conditions and to the optimization of anticancer drug combinations[114–116].

Equally important, drug target networks[117,118] that link approved or experimental drugs to their protein targets have helped researches to organize and visualize the considerable knowledge that exists about the interplay between diseases and drugs. Its analysis demonstrated that many drugs are palliative — that is, they do not target the actual disease-associated proteins but proteins in their network neighbourhood[117].

Disease classification. Contemporary approaches to the classification of human disease are based on observational correlations between pathological analysis of the patient and existing knowledge of clinical syndromes. Modern molecular diagnostic tools have shown the shortcomings of this methodology, reflecting both a lack of sensitivity in identifying preclinical disease and a lack of specificity in defining disease unequivocally. For example, hypertrophic cardiomyopathy, an inherited form of heart failure, is caused by a number of mutations in various sarcomeric proteins; however, the clinical phenotype, as well as the anatomical and functional pathophenotypes (assessed by echocardiographic assessment), are essentially indistinguishable from one another[119,120].

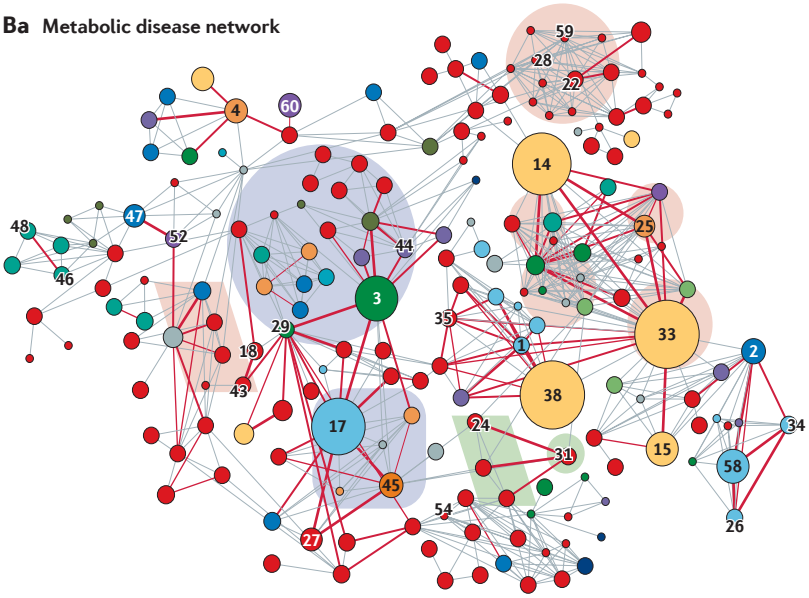
Current disease classification also tends to neglect the interconnected nature of many diseases. This failure is partly a response to the focused nature of medical training, as well as the reductionist paradigm that has driven medical diagnosis in the modern era. In an effort to correct this shortcoming, we recently proposed a systems-based network framework for defining human disease[121]. In this paradigm, the clinical pathophenotype is the systems-driven consequence of a series of linked networks that incorporate several components. First, there is the primary disease-causing gene, which contains a mutation — for example, the sickle cell mutation in the sixth amino acid of the haemoglobin β-chain. The mutated gene is a node in the interactome, and it interacts with a host of disease-modifying genes, including those that control intermediate pathophenotypes (or endopathophenotypes) that are common to all diseases and their network-based determinants. Intermediate pathophenotypes include inflammation and other immune responses, thrombosis or haemorrhage, fibrosis, aberrant cell proliferation, apoptosis or necrosis. In addition, there are environmental (and behavioural) determinants of disease, including both those that modulate gene expression at the transcriptional or epigenetic levels and those that cause post-translational modification of the proteome, and their influence on functional (protein, cell or organ) phenotype. These subnetwork determinants of the disease together give rise to clinical phenotypes that are highly individual, not only in complex diseases but also in 'simple' Mendelian conditions[121]. This framework exposes the challenges of a network-based disease classification: many of the factors affecting the disease module remain unknown or poorly mapped. Still, a network-based disease classification not only uncovers the gaps in our experimental and theoretical knowledge, but also demonstrates that only an integrated programme has the potential to provide a useful framework, by defining disease susceptibility, predicting disease outcome and identifying tailored therapeutic strategies.

Aa Human disease network

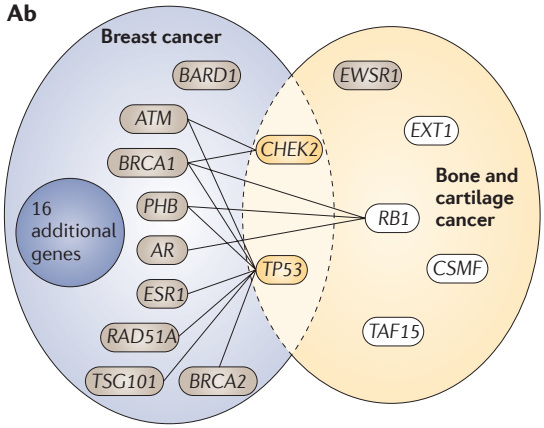


1. Aldosteronism
2. Alzheimer's disease
3. Anaemia, congenital deserythropoietic
4. Asthma
5. Ataxia-telangiectasia
6. Atherosclerosis
7. Blood group
8. Breast cancer
9. Cardiomyopathy
10. Cataract
11. Charcot–Marie–Tooth disease
12. Colon cancer
13. Complement component deficiency
14. Coronary artery disease
15. Coronary spasm
16. Deafness
17. Diabetes mellitus
18. Enolase-β deficiency
19. Epidermolysis bullosa
20. Epilepsy
21. Fanconi's anaemia
22. Fatty liver
23. Gastric cancer
24. Gilbert's syndrome
25. Glaucoma 1A
26. Goitre congenital
27. HARP syndrome
28. HELLP syndrome
29. Haemolytic anaemia
30. Hirschprung disease
31. Hyperbilirubinaemia
32. Hypertension
33. Hypertension diastolic
34. Hyperthyroidism
35. Hypoaldosteronism
36. Leigh syndrome
37. Leukaemia
38. Low renin hypertension
39. Lymphoma
40. Mental retardation
41. Muscular dystrophy
42. Myocardial infarction
43. Myopathy
44. Nucleoside phosphorylase deficiency
45. Obesity
46. Paraganglioma
47. Parkinson's disease
48. Pheochromocytoma
49. Prostate cancer
50. Pseudohypoaldosteronism
51. Retinitis pigmentosa
52. Schizoaffective disorder
53. Spherocytosis
54. Spina bifida
55. Spinocerebellar ataxia
56. Stroke
57. Thyroid carcinoma
58. Total iodide organification defect
59. Trifunctional protein deficiency
60. Unipolar depression

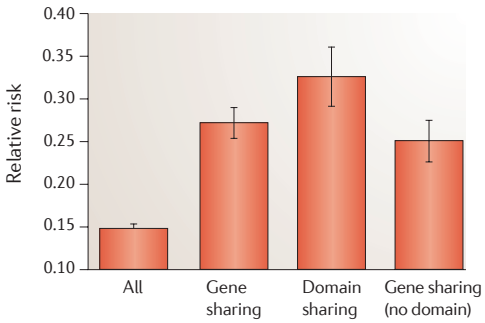
Ba Metabolic disease network



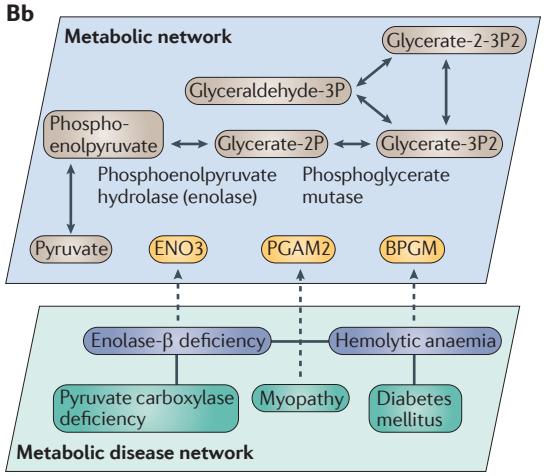
Ab



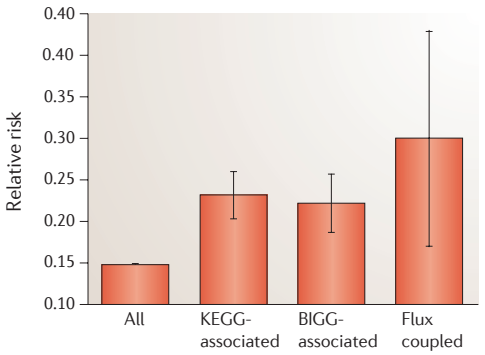
Ac Comorbidity



Bb



Bc Comorbidity



Conclusions

In summary, similar to an automotive technician's inability to fix a car's electrical problem without an accurate assembly and wiring diagram, a comprehensive understanding of most diseases requires a map of the cell's intricate wiring diagram, the breakdown of which is ultimately responsible for the emergence of a particular disease phenotype. Network medicine seeks to offer this understanding.

Yet, progress towards a reliable network-based approach to disease is currently limited by the incompleteness of the available interactome maps and the limitations of the existing tools to explore the role of networks in disease. For example, investigators are forced to apply traditional statistical tools to network data, assuming that the quantities of interest follow a normal distribution. However, they do not — everything from degree distributions to metabolite concentrations are known to be 'fat tailed'. Another assumption of current statistical tools is that quantities characterizing various activity patterns (for example, molecular concentrations and expression patterns) are independent variables. Again, they are not — most activity patterns in the cell are correlated. Thus, there is a real need to develop statistical tools that are reliable in the context of the interconnected environment of the cell. Lastly, although some principles that are widely used in network medicine are well supported by experimental evidence, such as the local hypothesis (BOX 3), others, such as the parsimony principle or the overlap between topological, functional and disease modules, remain to be quantified and validated.

As helpful as analogies can be, we must realize that there is a fundamental difference between the automotive technician and the physician: the technician can swap the broken component with one that functions correctly. This is a futuristic view of medicine — most drugs do not cure, but only alter the symptoms and signs of the disease. However, an integrated understanding of the interactions among the genome, the proteome, the environment and the pathophenome, mediated by the underlying cellular network, offers a basis for future advances. Such advances will help us to understand the structure and the workings of the wiring diagram — the prerequisite towards identifying the components whose functions need to be maintained and those whose activity must be altered with drugs. To summarize, in order to generate the local interventions that may cure a particular disease, we cannot avoid understanding the cells' global organization — the 'think globally, act locally' paradigm of network medicine.

1. Zhao, Y. & Jensen, O. N. Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichments techniques. *Proteomics* **9**, 4632–4641 (2009).
2. Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nature Methods* **6**, 83–90 (2008).
3. Goldstein, D. B. Common genetic variation and human traits. *N. Engl. J. Med.* **360**, 1696–1698 (2009).
4. Schadt, E. E. Molecular networks as sensors and drivers of common human diseases. *Nature* **461**, 218–223 (2009).
5. Barabási, A.-L. Network medicine — from obesity to the "diseasome". *N. Engl. J. Med.* **357**, 404–407 (2007).
6. Pawson, T. & Linding, R. Network medicine. *FEBS Lett.* **582**, 1266–1270 (2008).
7. Zanzoni, A., Soler-López, M. & Aloy, P. A network medicine approach to human disease. *FEBS Lett.* **583**, 1759–1765 (2009).
8. Ideker, T. & Sharan, R. Protein networks in disease. *Genome Res.* **18**, 644–652 (2008).
9. Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437**, 1173–1178 (2005).
10. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
11. Jeong, H. *et al.* The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
12. Fell, D. A. & Wagner, A. The small world of metabolism. *Nature Biotech.* **18**, 1121–1122 (2000).
13. Duarte, N. C. *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl Acad. Sci. USA* **104**, 1777–1782 (2007).
14. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
15. Linding, R. *et al.* NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res.* **36**, D695–D699 (2008).
16. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell* **120**, 15–20 (2005).
17. Reynolds, A. *et al.* Rational siRNA design for RNA interference. *Nature Biotech.* **22**, 326–330 (2004).
18. Stuart, J. M. *et al.* A Gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
19. Boone, C., Bussey, H. & Andrews, B. J. Exploring genetic interactions and networks with yeast. *Nature Rev. Genet.* **8**, 437–449 (2007).
20. Beltrao, P., Cagney, G. & Krogan, N. Quantitative genetic interactions reveal biological modularity. *Cell* **141**, 739–745 (2010).
21. Reverter, A., Ingham, A. & Dalrymple, B. P. Mining tissue specificity, gene connectivity and disease association to reveal a set of genes that modify the action of disease causing genes. *BioData Min.* **1**, 8 (2008).
22. Lage, K. *et al.* A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl Acad. Sci. USA* **105**, 20870–20875 (2008).
23. Lage, K. *et al.* Dissecting spatio-temporal protein networks driving human heart development and related disorders. *Mol. Syst. Biol.* **6**, 381 (2010).
24. Schwartz, A. S., Yu, J., Gardenour, K. R., Finley, R. L. & Ideker, T. Cost-effective strategies for completing the interactome. *Nature Methods* **6**, 55–61 (2009).
25. Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110 (2008).
26. Kirouac, D. *et al.* Dynamic interaction networks in a hierarchically organized tissue. *Mol. Syst. Biol.* **6**, 417 (2010).
27. Barabási, A.-L. & Oltvai, Z. Network biology: understanding the cell's functional organization. *Nature Rev. Genet.* **5**, 101–113 (2004).
28. Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).
29. Zhu, X., Gerstein, M. & Snyder, M. Getting connected: analysis and principles of biological networks. *Genes Dev.* **21**, 1010–1024 (2007).
30. Caldarelli, G. *Scale Free Networks* (Oxford Univ. Press, UK, 2007).
31. Albert, R. Scale-free networks in cell biology. *J. Cell Sci.* **118**, 4947–4957 (2005).
32. Newman, M., Barabási, A.-L. & Watts, D. J. *The Structure and Dynamics of Networks* (Princeton Univ. Press, USA, 2006).
33. Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM®). *Nucleic Acids Res.* **37**, D793–D796 (2009).
34. Jeong, H. *et al.* Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
35. Fraser, H. B. *et al.* Evolutionary rate in the protein interaction network. *Science* **296**, 750–752 (2002).
36. Eisenberg, E. & Levanon, E. Y. Preferential attachment in the protein network evolution. *Phys. Rev. Lett.* **91**, 138701 (2003).
37. Saeed, R. & Deane, C. M. Protein protein interactions, evolutionary rate, abundance and age. *BMC Bioinformatics* **7**, 128 (2006).
38. Jordan, I. K., Wolf, Y. I. & Koonin, E. V. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evol. Biol.* **3**, 5 (2003).
39. Wachi, S., Yoneda, K. & Wu, R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* **21**, 4205–4208 (2005).
40. Jonsson, P. F. & Bates, P. A. Global topological features of cancer proteins in the human interactome. *Bioinformatics* **22**, 2291–2297 (2006).
41. Xu, J. & Li, Y. Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics* **22**, 2800–2805 (2006).
This paper shows that disease genes can be discovered by exploiting the topological features of the protein–protein interaction network.
42. Goh, K.-I. *et al.* The human disease network. *Proc. Natl Acad. Sci. USA* **104**, 8685–8690 (2007).
This paper builds the first disease network by linking diseases that share disease genes, and it shows that most disease genes are non-essential and are not encoded by hub proteins.
43. Feldman, I., Rzhetsky, A. & Vitkup, D. Network properties of genes harboring inherited disease mutations. *Proc. Natl Acad. Sci. USA* **105**, 4323–4328 (2008).
44. Hartwell, L. H., Hopfield, J. J. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).
45. Oti, M. *et al.* Predicting disease genes using protein-protein interactions. *J. Med. Genet.* **43**, 691–698 (2006).
This paper explores the degree to which proteins linked to known disease genes are also associated with the same phenotype.
46. Gandhi, T. *et al.* Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genet.* **38**, 285–293 (2006).
47. Girvan, M. & Newman, M. E. Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* **99**, 7821–7826 (2002).

48. Palla, G., Derényi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818 (2005).
49. Ahn, Y.-Y., Bagrow, J. P. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466**, 761–764 (2010).
50. 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).
51. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A.-L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002).
52. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108–1113 (2007).
53. Hirschhorn, J. N. Genomewide association studies — illuminating biologic pathways. *N. Engl. J. Med.* **360**, 1699–1701 (2009).
54. Krauthammer, M. *et al.* Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc. Natl Acad. Sci. USA* **101**, 15148–15153 (2004).
55. Franke, L. *et al.* Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.* **78**, 1011–1025 (2006).
56. Iossifov, I., Zheng, T., Baron, M., Gilliam T. C. & Rzhetsky, A. Genetic-linkage mapping of complex hereditary disorders to a whole-genome molecular-interaction network. *Genome Res.* **18**, 1150–1162 (2008).
57. Navlakha, S. & Kingsford, C. The power of protein interaction networks for associating genes with diseases. *Bioinformatics* **26**, 1057–1063 (2010).
This paper compares the available disease gene prediction methods, showing that random walk-based tools outperform clustering- and linkage-based approaches.
58. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotech.* **25**, 309–316 (2007).
59. Lee, E. *et al.* Analysis of AML genes in dysregulated molecular networks. *BMC Bioinformatics* **10**, S2 (2009).
60. Bonifaci, N. *et al.* Biological processes, properties and molecular wiring diagrams of candidate low-penetrance breast cancer susceptibility genes. *BMC Med. Genomics* **1**, 62 (2008).
61. Heiser, L. M. *et al.* Integrated analysis of breast cancer cell lines reveals unique signaling pathways. *Genome Biol.* **10**, R31 (2009).
62. Chuang, H.-Y. *et al.* Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3**, 140 (2007).
63. Nibbe, R. K. *et al.* Discovery and scoring of protein interaction subnetworks discriminative of late stage human colon cancer. *Mol. Cell. Proteomics* **8**, 827–845 (2009).
64. Chang, W. *et al.* Identification of novel hub genes associated with liver metastasis of gastric cancer. *Int. J. Cancer* **125**, 2844–2853 (2009).
65. Ergün, A., Lawrence, C. A., Kohanski, M. A., Brennan, T. A. & Collins, J. J. A network biology approach to prostate cancer. *Mol. Syst. Biol.* **3**, 82 (2007).
66. Taylor, I. W. *et al.* Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotech.* **27**, 199–204 (2009).
This paper examines whether the modular nature of the hubs can be used to predict patient outcome, with applications to breast cancer.
67. Moran, L. B. & Graeber, M. B. Towards a pathway definition of Parkinson's disease: a complex disorder with links to cancer, diabetes and inflammation. *Neurogenetics* **9**, 1–13 (2008).
68. Ray, M., Ruan, J. & Zhang, W. Variations in the transcriptome of Alzheimer's disease reveal molecular networks involved in cardiovascular diseases. *Genome Biol.* **9**, R148 (2008).
69. Hwang, D. *et al.* A systems approach to prion disease. *Mol. Syst. Biol.* **5**, 252 (2009).
70. Wheelock, C. E. *et al.* Systems biology approaches and pathway tools for investigating cardiovascular disease. *Mol. Biosyst.* **5**, 588–602 (2009).
71. Calvano, S. E. *et al.* A network-based analysis of systemic inflammation. *Nature* **437**, 1032–1037 (2005).
72. Iliopoulos, D. *et al.* Integrative microRNA and proteomic approaches identify novel osteoarthritis genes and their collaborative metabolic and inflammatory networks. *PLoS ONE* **3**, e3740 (2008).
73. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**, 429–435 (2008).
74. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428 (2008).
75. Dobrin, R. *et al.* Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome Biol.* **10**, R55 (2009).
76. Hwang, S. *et al.* A protein interaction network associated with asthma. *J. Theor. Biol.* **252**, 722–731 (2008).
77. Liu, M. *et al.* Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genet.* **3**, e96 (2007).
78. Presson, A. P. *et al.* Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome. *BMC Syst. Biol.* **2**, 95 (2008).
79. Uetz, P. *et al.* Herpesviral protein networks and their interaction with the human proteome. *Science* **311**, 239–242 (2006).
80. Calderwood, M. A. *et al.* Epstein–Barr virus and virus human protein interaction maps. *Proc. Natl Acad. Sci. USA* **104**, 7606–7611 (2007).
81. Bordbar, A., Lewis, N. E., Schellenberger, J., Palsson, B. Ø. & Jamshidi, N. Insight into human alveolar macrophage and *M. tuberculosis* interactions via metabolic reconstructions. *Mol. Syst. Biol.* **6**, 422 (2010).
82. Turnbaugh, P. J. & Gordon, J. I. An invitation to the marriage of metagenomics and metabolomics. *Cell* **134**, 708–713 (2008).
83. Goehler, H. *et al.* A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Mol. Cell* **15**, 853–865 (2004).
84. Lim, J. *et al.* A Protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* **125**, 801–814 (2006).
This paper used yeast two-hybrid assays to map the interactions of spinocerebellar ataxia proteins with other human proteins to build the ataxia disease module.
85. Pujana, M. A. *et al.* Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nature Genetics* **39**, 1338–1349 (2007).
86. Camargo, L. M. *et al.* Disrupted in Schizophrenia 1 Interactome: evidence for the close connectivity of risk genes and a potential synaptic basis for schizophrenia. *Mol. Psychiatry* **12**, 74–86 (2007).
87. Amino, T. *et al.* Redefining the disease locus of 16q22.1-linked autosomal dominant cerebellar ataxia. *J. Hum. Genet.* **52**, 643–649 (2007).
88. Kohler, S. *et al.* Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* **82**, 949–958 (2008).
89. Vanunu, O. *et al.* Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.* **6**, e1000641 (2010).
90. Park, J. *et al.* The impact of cellular networks on disease comorbidity. *Mol. Syst. Biol.* **5**, 262 (2009).
This paper shows that diseases that share genes or that involve proteins that interact with each other show elevated comorbidity, demonstrating correlations between the structure of cellular networks and disease patterns in the population.
91. Dudley, A. M., Janse, D. M., Tanay, A., Shamir, R. & Church, G. M. A global view of pleiotropy and phenotypically derived gene function in yeast. *Mol. Syst. Biol.* **1**, 1 (2005).
92. Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.* **5**, 321 (2009).
93. Lee, D.-S. *et al.* The implications of human metabolic network topology for disease comorbidity. *Proc. Natl Acad. Sci. USA* **105**, 9880–9885 (2008).
This paper constructs a metabolic disease association by linking diseases associated with adjacent metabolic reactions and finding elevated comorbidity for the linked diseases.
94. Lu, M. *et al.* An analysis of human microRNA and disease associations. *PLoS ONE* **3**, e3420 (2008).
95. Rzhetsky, A. *et al.* Probing genetic overlap among complex human phenotypes. *Proc. Natl Acad. Sci. USA* **104**, 11694–11699 (2007).
This analysis of patient records indicated that disease phenotypes form a highly connected network of strong pairwise correlations, helping the researchers to estimate the size of putative genetic overlaps.
96. Hidalgo, C. *et al.* A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.* **5**, e1000353 (2009).
This paper introduced a PDN by linking diseases with significant comorbidity using data obtained from the disease history of 30 million Medicare patients. From this, the researchers built an open-access comorbidity database.
97. van Driel, M. A. *et al.* A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.* **14**, 535–542 (2006).
98. Suthram, S. *et al.* Network-based elucidation of human disease network-based elucidation of human disease enriched for pluripotent drug targets. *PLoS Comput. Biol.* **6**, e1000662 (2010).
99. Liu, Y. I., Wise, P. H. & Butte, A. J. The "etiome": identification and clustering of human disease etiological factors. *BMC Bioinformatics* **10**, S14 (2009).
100. Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J. & Bork, P. Drug target identification using side-effect similarity. *Science* **321**, 263–266 (2008).
101. Kuhn, M., Campillos, M, Letunic, I, Jensen, L. J. & Bork, P. A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.* **6**, 343 (2010).
102. Audouze, K. *et al.* Deciphering diseases and biological *targets.* for environmental chemicals using toxicogenomics networks. *PLoS Comput. Biol.* **6**, e1000788 (2010).
103. Schadt, E. E., Friend, S. H. & Shaywitz, D. A. A network view of disease and compound screening. *Nature Rev. Drug Disc.* **8**, 286–295 (2009).
104. Hopkins, A. L. Drug discovery: predicting promiscuity. *Nature* **462**, 167–168 (2009).
105. Chu, L. & Chen, B. S. Construction of a cancer-perturbed protein-protein interaction network for discovery of apoptosis drug targets. *BMC Syst. Biol.* **2**, 56 (2008).
106. Azmi, A., Wang, Z., Philip, P. A., Mohammad, R. M. & Sarkar, F. H. Proof of concept: a review on how network and systems biology approaches aid in the discovery of potent anticancer drug combinations. *Mol. Cancer Ther.* 1 Nov 2010 (doi:10.1158/1535-7163.MCT-10-0642).
107. Zhao, S. & Li, S. Network-based relating pharmacological and genomic spaces for drug target identification. *PLoS ONE* **5**, e11764 (2010).
108. Fong, S. S. & Palsson, B. Ø. Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nature Genet.* **36**, 1056–1058 (2004).
109. Segrè, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl Acad. Sci. USA* **99**, 15112–15117 (2002).
110. Shen, Y. *et al.* Blueprint for antimicrobial hit discovery targeting metabolic networks. *Proc. Natl Acad. Sci. USA* **107**, 1082–1087 (2010).
111. Motter, A. E., Gulbahce, N., Almaas, E. & Barabási, A.-L. Predicting synthetic rescues in metabolic networks. *Mol. Syst. Biol.* **4**, 168 (2008).
112. Nolan, G. P. What's wrong with drug screening today. *Nature Chem. Biol.* **3**, 187–191 (2007).
113. Csermely, P., Agoston, V. & Pongor, S. The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol. Sci.* **26**, 178–182 (2005).
114. Motter, A. E. Improved network performance via antagonism: from synthetic rescues to multi-drug combinations. *Bioessays* **32**, 236–245 (2010).
115. Yang, K. *et al.* Finding multiple target optimal intervention in disease related molecular network. *Mol. Syst. Biol.* **4**, 228 (2008).
116. Vazquez, A. Optimal drug combinations and minimal hitting sets. *BMC Syst. Biol.* **3**, 81 (2009).
117. Yildirim, M. A. *et al.* Drug–target network. *Nature Biotech.* **25**, 1119–1126 (2007).
118. Keiser, M. J. *et al.* Predicting new molecular targets for known drugs. *Nature* **462**, 175–181 (2009).
119. Ho, C. Y. & Seidman, C. E. A contemporary approach to hypertrophic cardiomyopathy. *Circulation* **113**, e858–e862 (2006).
120. Morita, Y. *et al.* Shared genetic causes of cadiac hypertrophy in children and adults. *N. Engl. J. Med.* **358**, 1899–1908 (2008).
121. Loscalzo, J., Kohane, I., Barabási, A.-L. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Mol. Syst. Biol.* **3**, 124 (2007).

122. Dreze, M. *et al.* High-quality binary interactome mapping. *Meth. Enzymol.* **470**, 281–315 (2010).
123. Ewing, R. M. *et al.* Large-scale mapping of human protein–protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3**, 89 (2007).
124. Cusick, M. E. *et al.* Literature-curated protein interaction datasets. *Nature Methods* **6**, 39–46 (2009).
125. Ma, H. *et al.* The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol. Syst. Biol.* **3**, 135 (2007).
126. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
127. Han, J. D. *et al.* Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88–93 (2004).
128. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
129. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827 (2002).
130. Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. & Gerstein, M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.* **3**, e59 (2007).
131. Wu, X., Jiang, R., Zhang, M. Q. & Li, S. Network-based global inference of human disease genes. *Mol. Syst. Biol.* **4**, 189 (2008).

Acknowledgements
We thank Z. Oltvai, A. Sharma, D.-S. Lee and J. Park for useful discussions and suggestions. A.L.B. and N.G. were supported by the US National Institutes of Health (NIH) through the Center of Excellence in Genomic Sciences (CEGS), and J.L. was supported by NIH grants HL061795 (Merit Award), HL81587, HL70819 and HL48743.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION

Albert-László Barabási's homepage: http://www.barabasilab.com
B-cell interactome (BCI): http://amdec-bioinfo.cu-genome.org/html/BCellInteractome.html
Biochemical Genetic and Genomics knowledgebase (BIGG): http://bigg.ucsd.edu
Biological General Repository for Interaction Datasets (BioGRID): http://thebiogrid.org
Biomolecular Interaction Network Database (BIND): http://www.bind.ca
CBS prediction server: http://www.cbs.dtu.dk/services
Database of Interacting Proteins (DIP): http://dip.doe-mbi.ucla.edu/dip
Human Protein Reference Database (HPRD): http://www.hprd.org
JASPAR: http://jaspar.genereg.net
Kyoto Encyclopedia of Genes and Genomes (KEGG): http://www.genome.jp/kegg
microRNA: http://www.microrna.org
miRBase: http://www.mirbase.org
miRDB: http://mirdb.org
miRecords: http://mirecords.biolead.org
Molecular Interaction database (MINT): http://mint.bio.uniroma2.it/mint
Munich Information Center for Protein Sequence (MIPS)
Mammalian Protein–Protein Interaction Database: http://mips.helmholtz-muenchen.de/proj/ppi
NetPhorest: http://netphorest.info/index.php
Phospho.ELM: http://phospho.elm.eu.org
Phosphorylation site database (PHOSIDA): http://www.phosida.com
PhosphoSite: http://www.phosphosite.org
PicTar: http://pictar.mdc-berlin.de
Protein Interaction database (IntAct): http://www.ebi.ac.uk/intact
STRING: http://string-db.org
TarBase: http://diana.cslab.ece.ntua.gr/tarbase
TargetScan: http://www.targetscan.org
TRANSFAC: http://www.gene-regulation.com/pub/databases.html
Universal Protein Binding Microarray Resource for Oligonucleotide Binding Evaluation (UniPROBE): http://the_brain.bwh.harvard.edu/uniprobe

ALL LINKS ARE ACTIVE IN THE ONLINE PDF