Review

Human diseases through the lens of network biology

Laura I. Furlong

Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute (IMIM), Universitat Pompeu Fabra (UPF), C/Dr. Aiguader, 88, 08003 – Barcelona, Spain

One of the challenges raised by next generation sequencing (NGS) is the identification of clinically relevant mutations among all the genetic variation found in an individual. Network biology has emerged as an integrative and systems-level approach for the interpretation of genome data in the context of health and disease. Network biology can provide insightful models for genetic phenomena such as penetrance, epistasis, and modes of inheritance, all of which are integral aspects of Mendelian and complex diseases. Moreover, it can shed light on disease mechanisms via the identification of modules perturbed in those diseases. Current challenges include understanding disease as a result of the interplay between environmental and genetic perturbations and assessing the impact of personal sequence variations in the context of networks. Full realization of the potential of personal genomics will benefit from network biology approaches that aim to uncover the mechanisms underlying disease pathogenesis, identify new biomarkers, and guide personalized therapeutic interventions.

Using networks to understand disease

During the past decade, the study of genetic diseases has been revolutionized by the application of high-throughput technologies and computational approaches. Although these methodologies were first employed to find the genetic determinants of complex diseases (see Glossary), as exemplified by genome-wide association studies (GWAS) [1,2], NGS has recently been used to identify the gene variants responsible for both Mendelian [3] and complex disorders [4–6]. As a result, we have an impressive amount of data on sequence alterations and biomolecular profiles (mRNA expression, miRNA and noncoding RNA profiling, proteomics, and metabolomics measurements) for many human diseases, which can be accessed from specialized databases and publications (for a recent review see [7]). However, we still have not succeeded in translating this wealth of information into actionable knowledge about disease pathogenesis for the development of better strategies for disease prevention, diagnosis, and treatment. Progress is limited by the difficulties in assessing the functional consequences of disease-associated sequence variants [8] and understanding how phenotype is affected by the combined effect of environmental and genomic variation [9].

Biological network analysis is one approach to distill these large data sets into clinically actionable knowledge for disease diagnosis, prognosis, and treatment (Table 1). This is predicated on the idea that diseases are a consequence of perturbations in biological networks [10], which is rooted in seminal work on model organisms (e.g., [11]). Networks serve as a paradigm for data integration and analysis, providing a systems-level understanding of the mechanisms underlying diseases. Protein interaction networks (PINs) in particular have become a valuable resource in this context [10]. PINs are derived from high-throughput approaches, including yeast two-hybrid screens, immunoprecipitation studies followed by mass spectrometry analysis, and small-scale experiments [12]. The current estimates suggest that the human interactome comprises approximately 130 000–650 000 protein interactions [13,14]; however, only a subset of these has been experimentally identified.

Networks have been used to gain insight into disease mechanisms [15–21], study comorbidities [22–24], analyze therapeutic drugs and their targets [25,26], and discover novel network-based biomarkers [27]. These early suc-

Glossary

Allelic heterogeneity: reflects different allelic mutations at the same locus associated with multiple disorders.

Epistasis: also known as synthetic or synergistic interaction between genes, in which the contribution of a gene to a phenotype depends on its interaction with several other genes.

Expressivity: measures the extent to which a given genotype is expressed at the phenotypic level, due to variation in the genomic background or the effect of environmental factors [84].

Genetic or locus heterogeneity: a disease with high locus heterogeneity has several variants at the population level. Locus heterogeneity is common in syndromes resulting from failure of a complex pathway (e.g., Usher syndrome) [85].

Mendelian and complex diseases: whereas Mendelian disorders are the result of alterations in one or several genes that can be traced in family pedigrees, complex diseases arise as a consequence of the combined effect of multiple genetic determinants, which may vary between individuals, and environmental factors. The term 'complex disease' is also used to refer to those diseases where the phenotype cannot be easily predicted from the genotype. This is proposed to be due to the interaction between genes (epistasis) [86], modulation by environmental factors or stochastic processes [73], or epigenetic changes [87]. It is important to note that some of these factors, such as incomplete penetrance and variable expressivity, can also make it difficult to predict the phenotype of Mendelian disorders based solely on genotype.

Penetrance: the probability of an individual to manifest a change in the genotype. This probability is a function of the presence of modifiers, epistatic genes, and the environment [85].

Protein interaction network (PIN): a schematic mapping proteins (nodes) to other proteins, where the lines connecting two proteins (edges) represent a physical binding of the two proteins or a predicted interaction between them.

Corresponding author: Furlong, L.I. (lfurlong@imim.es).

Keywords: systems biology; protein interaction networks; disease; personal genomes; translational bioinformatics; network biology.

0168-9525/$ – see front matter http://dx.doi.org/10.1016/j.tig.2012.11.004

Table 1. Tools for network analysis and visualization

Name	URL	Refs
Web resources for PINs and pathways		
Pathguide	http://www.pathguide.org/	[88]
HIPPIE	http://cbdm.mdc-berlin.de/tools/hippie/information.php	[89]
BioGRID	http://thebiogrid.org/	[90]
Intact	http://www.ebi.ac.uk/intact/	[91]
STRING	http://string-db.org/	[92]
MINT	http://mint.bio.uniroma2.it/mint/Welcome.do	[93]
HPRD	http://www.hprd.org/	[94]
Network analysis and visualization		
Visant	http://visant.bu.edu/	[95]
Cytoscape	http://www.cytoscape.org/	[96]
CellDesigner	http://www.celldesigner.org/	[97]
Gephi	http://gephi.org/	[98]
Graphviz	http://www.graphviz.org/	[99]
Ondex	http://www.ondex.org/	[100]
Osprey	http://biodata.mshri.on.ca/osprey/servlet/Index	[101]
Biotapestry	http://www.biotapestry.org/	[102]
Patika	http://www.cs.bilkent.edu.tr/~patikaweb/	[103]
Biolayout Express3D	http://www.biolayout.org/	[104]
Arena3D	http://arena3d.org/	[105]
BiologicalNetworks	http://biologicalnetworks.net/	[106]

cesses indicate that network biology can shed light on the complex relation between genotype and phenotype in human diseases. Here, I review recent literature on network analysis related to disease, with particular emphasis on topological studies of PINs, and discuss some of the open questions in human disease research. Related topics, such as reverse engineering gene regulatory networks, dynamic network modeling, and network-based prediction of disease genes, will not be covered here. Interested readers can find information on these topics elsewhere (see, for example, [28–30]).

Disease proteins: hubs, bottlenecks, or peripheral nodes?

One of the simplest ways that network analysis can provide insight into human disease is to assess the network properties of genes underlying the disease, which might reveal important clues about its etiology. This is based on the assumption that there is a tight relation between network structure and biological function. For example, in the yeast *Saccharomyces cerevisiae*, proteins with a high degree (Box 1) in a PIN are more likely to be encoded by essential genes [31,32]. Thus, mutations that affect hubs are expected to perturb the network, whereas those at the periphery have less effect [33]. Following this line of reasoning, several early studies proposed that disease genes encode hubs in PINs [27,34–36]. However, a systematic analysis of Mendelian diseases did not clearly support this idea [15]. The authors argued that, to assess the node degree of disease proteins, it was important to distinguish between essential and nonessential proteins. In their analysis, essential proteins were more frequently associated with hubs, whereas disease genes that were not essential did not encode hubs. They concluded that nonessential disease genes occupy functionally peripheral and topologically neutral positions in the cellular network [15].

Recently, these results have been reevaluated in the context of the argument that the network properties of disease genes differ for genes implicated in complex or Mendelian diseases. For instance, one group [37] classified genes as Mendelian (M), complex (C), or Mendelian and complex (MC), and analyzed their properties in a PIN. They found that C genes encode proteins with a similar degree as proteins encoded by MC genes, and that the degree is significantly higher than that of proteins encoded by M genes. The degree of disease proteins was nevertheless smaller than the degree of proteins encoded by essential genes in the PIN. These findings indicate that proteins encoded by genes involved in both complex and Mendelian disorders have more interacting partners than do proteins encoded by Mendelian disorder genes alone. Similar findings were reported by another group [38], but not in a third study [39], in which there was no difference found between the degree of Mendelian and nondisease genes.

These discrepancies might originate from methodological issues. The definition of hub varies with study (e.g., 20% of the nodes in a network with the highest degree are 'hubs') [10] and is highly dependent on the data set. Thus, the contradictory results might also be due to the variability of the different interactome data sets currently available [40,41]. A recent analysis of the most popular protein interaction databases revealed considerable differences in the interaction partners between these databases [41]. However, the disagreements might indicate that network properties of disease genes do not differ for genes implicated in different types of disease; some authors propose that there is no sharp separation between complex and Mendelian diseases [2] and, thus, the properties of complex and Mendelian disease genes would be similar in PINs.

In addition to the degree of connectedness, other network properties can be used to assess the importance of a given protein in the network, such as betweenness centrality and current information flow (Box 1). Both

Box 1. A brief guide to network analysis

The principal attributes of nodes in a network are described using the PIN around integrin beta-3 (ITGB3) as an example (Figure I).

Degree

k_i = number of edges of node i [I]

In Figure Ia, tallin-1 (TLN1) is connected to four proteins in the network, thus its degree is 4 (Equation I).

Distance

d_{ij} = shortest path length between nodes i and j [II]

The shortest path connecting nodes i and j is the one in which the lowest number of nodes are traversed to connect them (Equation II). In Figure Ib, the distance between paxillin (PXN) and Calcium and integrin-binding protein 1 (CIB1) is 2.

Clustering coefficient

The clustering coefficient of a node is a measure of the degree of interconnectivity of its neighbors. It is calculated as the number of edges between neighbors of node i (b_i, depicted as pink unbroken lines, Figure Ic) divided by the number of all possible edges (pink unbroken and broken lines) between them, according to Equation III:

$$C_i = \frac{2b_i}{k_i(k_i - 1)} \quad [III]$$

It ranges from zero (for a node that is part of a loosely connected group) to one (for a node at the center of a fully connected cluster). The clustering coefficient of TLN1 is 5/6.

Betweenness centrality

Betweenness centrality measures the global importance of a node in communicating between pairs of nodes in the network, considering the shortest paths. If we take protein tyrosine kinase 2 (PTK2) as an example (Figure Id), which is located in the shortest path between nodes PXN–CIB1, and TLN1–cannabinoid receptor type 1 (CB1), betweenness centrality is 0.0128. It is calculated by computing the fraction of shortest paths passing through node i (Equation IV):

$$B_i = \sum_{j=1}^{n}\sum_{k=1}^{j-1}\frac{g_{jk}(i)}{g_{jk}}, \quad [IV]$$

where $g_{jk}(i)$ is the number of shortest paths from j to k through i (green lines) and g_{jk} is the total number of shortest paths between j and k (green and blue lines). Then it is normalized by dividing by the number of possible edges between all the nodes in the network (not including node i) (Equation V):

$$\frac{(n-1)(n-2)}{2}$$

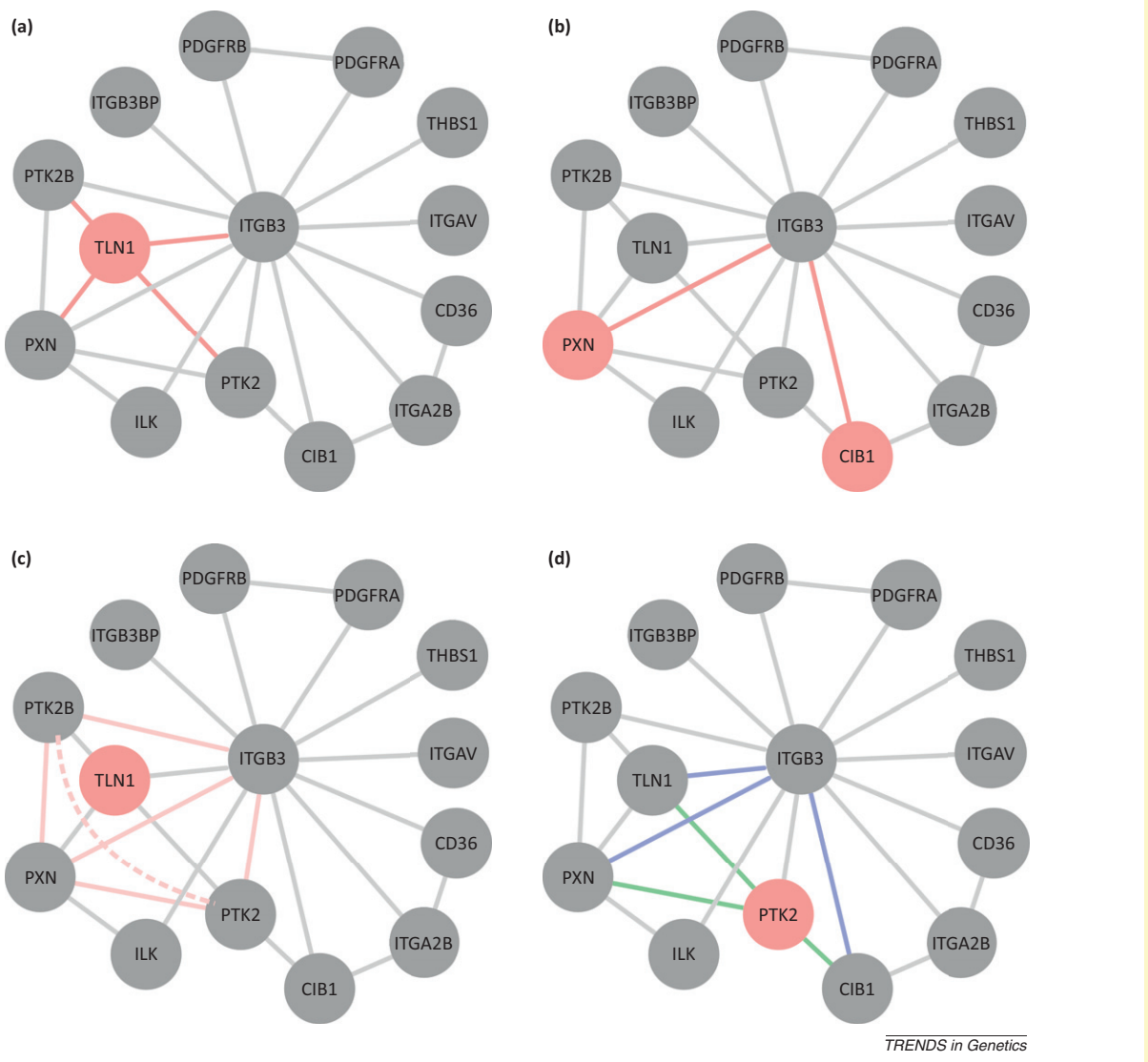

Figure I. The PIN around ITGB3 is used to illustrate some node attributes.

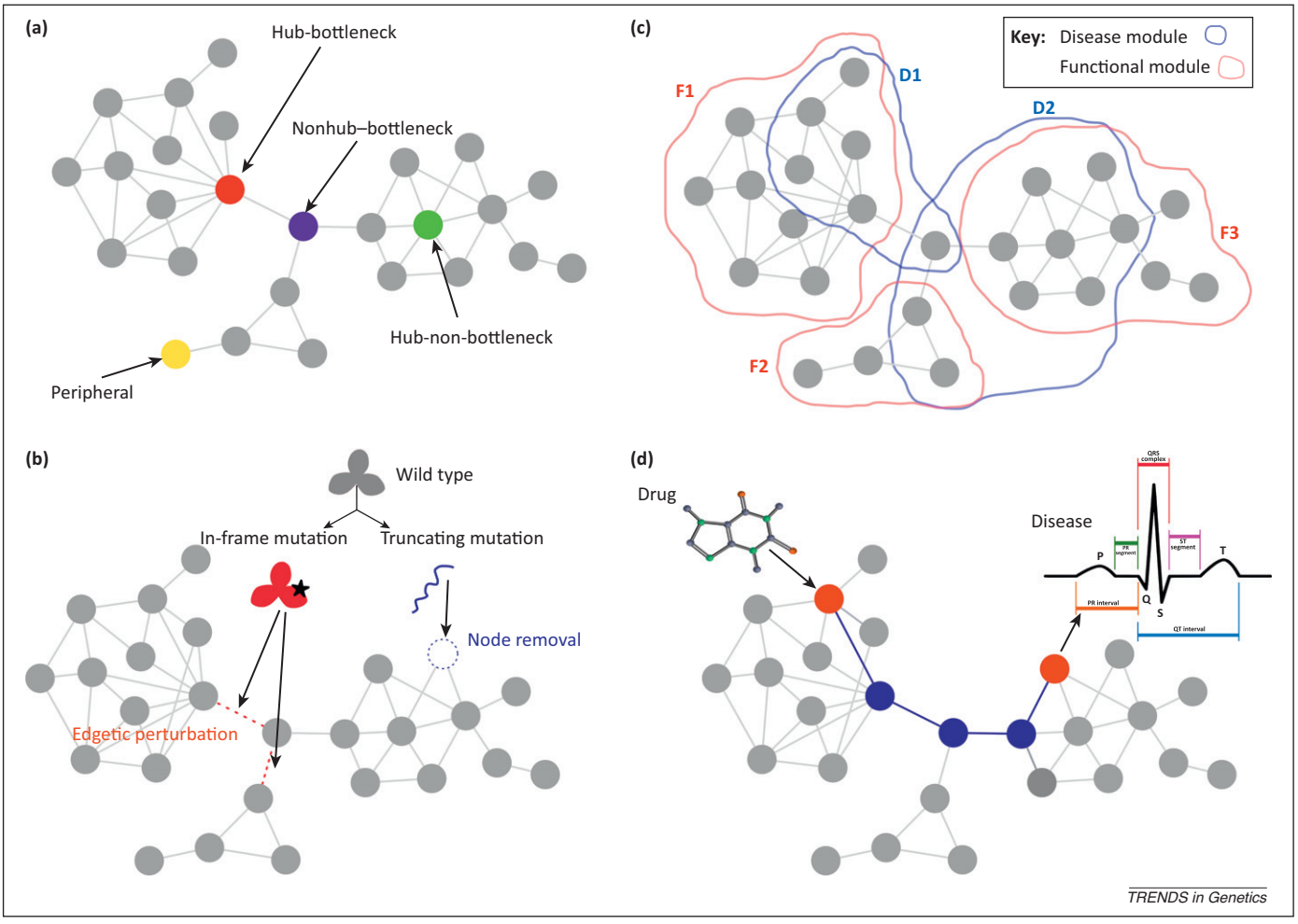

Figure 1. Different network attributes of disease genes are illustrated in a protein interaction network (PIN). **(a)** Nodes can be characterized by their role in the network, as peripheral, hubs, or bottlenecks. **(b)** The type of mutation of a protein (in-frame or truncating) leads to different perturbations in the PIN. Nonsense mutations, frame-shift mutations, and splice-site mutations lead to truncated proteins or completely destabilized proteins that can no longer maintain proper interactions with their protein partners. Such truncating mutations remove a node from the network and abolish all interactions. By contrast, subtle changes, such as single amino acid substitutions affecting binding sites or truncations that retain certain domains of the protein, may produce partially functional proteins that affect only a subset of their interaction partners. These in-frame mutations are referred to as edge-specific genetic, or edgetic, perturbations. **(c)** Different regions of a network can be identified according to the involvement of the nodes in common biological processes [functional (F) modules] or the same disorder [disease (D) modules]. Note that there is not a perfect overlap between functional and disease modules. In addition, different disease modules may overlap at least in part due to shared genetic determinants. **(d)** The interplay between environmental factors and genetic variation can be studied with network analysis. For example, the relation between the target of a drug and the protein associated with an adverse effect is modeled by calculating a distance measure in a PIN (edges highlighted in blue).

Mendelian and complex disease proteins have significantly higher betweenness centrality and current information flow in the interactome than do nondisease proteins [39]. Notably, the network properties of GWAS genes were not statistically different from nondisease genes, a finding that has implications for the likelihood of these being causal variants. These results suggest that proteins encoded by Mendelian and complex disease genes tend to occupy network positions that are central to the transmission of information through the network. Mendelian and complex disease proteins also have significantly lower clustering coefficients, suggesting that the number of connections between the neighbors of disease proteins is unusually low. The authors of the study proposed that proteins with these particular network properties can be thought of as 'broker' proteins [39]. Broker or bottlenecks proteins [42] are the sole connection to many other proteins, serving as bridges between different cellular functions. These proteins are in particularly fragile positions in the interactome (Figure 1a). Disruption of the function of these proteins would disturb the flow of information between cellular processes and lead to disease phenotypes. Functionally characterizing the different regions of the network that disease-associated bottleneck proteins connect could shed light on disease pathogenesis. Moreover, the proteins connecting disease-related modules may represent interesting drug targets. Notably, most current drugs do not target disease-associated proteins, but proteins located in their network neighborhood [25].

Network rewiring and disease

Another possible reason why no clear consensus about the network properties of disease proteins in PINs has emerged may be that the networks themselves are impacted by the disease state, confounding the role of particular proteins within the PIN. Thus, it is important to consider if the network itself is rewired in the disease state.

Network-perturbation models have been proposed to explain the molecular alterations observed in human Mendelian disorders [43]. In these models, changes in the

connectivity of the network give rise to different network topologies that underlie human diseases. Depending on the impact of the mutation at the protein level (truncating or in-frame; Figure 1b), the network of interactions in which the protein is immersed will be affected in different manners. This network-perturbation model was experimentally validated for two genes involved in autosomal recessive diseases and three genes involved in autosomal dominant diseases [43]. The analysis also suggested a relation between edgetic perturbations, treatment response, and disease severity. In addition, the model offered a network-based hypothesis to explain modes of inheritance. For 34 genes that are associated with both autosomal dominant and recessive diseases, they found that the fraction of in-frame mutations per gene was significantly higher than truncating mutations for dominant diseases. This finding suggests that the two types of mutation cause distinct perturbations in the network, leading to diseases with distinct modes of inheritance. The higher proportion of in-frame mutations in dominant diseases is in agreement with the current view of dominance as more likely produced by gain-of-function mutations [44]. The proposed model predicts that distinct edgetic perturbations in a protein might cause different disorders. Genes associated with multiple diseases that encode proteins with several domains were found to harbor in-frame mutations associated with different diseases distributed in distinct protein domains. Further experiments are needed to assess if these domain-specific, in-frame mutations lead to rewiring in the PIN.

A recent study went one step further and analyzed the distribution of in-frame and truncating mutations on protein interfaces using a structurally resolved interactome and information on gene mutations in Mendelian disorders [45]. Disease-associated in-frame mutations were found to be significantly enriched in the interaction interfaces of the proteins relative to the whole protein surface, but not in noninteracting domains. This result suggests that specific alterations of the interaction between two proteins caused by disease-associated mutations play a role in the pathogenesis of the disease. Moreover, the authors observed that truncating mutations did not show this distribution and were instead found randomly distributed throughout the protein. Furthermore, they explored the locus heterogeneity of diseases by assessing the distribution of in-frame mutations on two different proteins that cause the same disorder. The authors found that in-frame mutations were more likely to cause the same disorder than random expectation, and that these mutations were localized to protein interaction interfaces.

Together, these studies point to an important role for mutations that alter the connections in a PIN in several Mendelian diseases. To extend these observations to other disease-associated proteins, progress in protein structure prediction is required to increase the coverage of the structurally resolved PIN [45]. This will provide a means to determine whether similar mechanisms underlie complex disorders.

A novel family of methods collectively termed 'differential network mapping' could also be used to study network rewiring in disease [46]. These methods can experimentally map networks across multiple conditions or time points and, as such, are able to provide a detailed map of network rewiring in response to different cues. A key point of differential network mapping is that it identifies as the most interesting interactions those that change between the two conditions studied. For instance, by comparing healthy versus disease state networks, it would be possible to identify edgetic perturbations (e.g., loss or gain of edges between proteins) resulting from disease-associated genetic and environmental perturbations. Thus, differential network mapping is a promising approach to investigate the network dynamics during disease development and progression, as well as to monitor the network changes upon therapeutic interventions.

Modularity of human diseases

A backbone of network biology is the 'local hypothesis', which states that proteins involved in the same disease have a tendency to interact with each other, forming 'disease modules' [10]. It has been proposed that the study of the modularity of human disease genes will help in understanding disease pathogenesis, explain penetrance and expressivity [47], provide clues for the identification of therapeutic targets [48,49], and identify or prioritize new disease genes [50].

Disease modules are defined as a group of network components that contribute to a cellular function and, when disrupted, lead to disease [51]. Functional modules are groups of nodes in a network neighborhood with similar function. By contrast, a topological module is a particularly dense area of a network in which the nodes have a higher proportion of links between the components of the module than to components outside that module, but there is no constraint on the function of the nodes to define the module. It is generally assumed in the network biology literature that topological modules overlap with functional modules and with disease modules [10]. The concept of modularity of human diseases has been useful in the study of inherited ataxias [52], pancreatic cancer [20], Fanconi anemia [53], Walker–Warbur syndrome [54], and other Mendelian diseases [55], but until recently there was still limited evidence for its applicability to complex traits [47].

A recent study performed a large-scale analysis using topological and functional analysis of gene–disease association networks to assess the concept of modularity of Mendelian, complex, and environmental diseases [17]. Modularity was measured as the proportion of disease genes belonging to the same biological process, pathway, or PIN, and it was found to be a function of the locus heterogeneity of the disease. Thus, in diseases with low locus heterogeneity (no more than five associated genes), 75% of the genes belong to the same biological pathway. By contrast, for diseases with a higher locus heterogeneity (more than five associated genes), the modularity was lower (42%). Similar results were obtained with clusters of genes grouped by their shared diseases [17]. These results suggest that most human diseases are associated not with a single pathway but with a set of biological pathways. This is consistent with observations from different cancer types, such as pancreatic cancer [20], glioblastoma [56], and familial breast cancer [57], wherein the

disorder arises from different mutations in any one of multiple genes, but all of them encode proteins involved in related pathways. More importantly, these findings highlight that the overlap between disease modules and functional modules is not perfect (Figure 1c), in part because a disease module may comprise multiple functional modules in the cell.

The notion that disease proteins are bottleneck nodes in PINs [39] is consistent with this module definition; the involvement of several pathways suggest that crosstalk between them plays an important role in disease pathogenesis [58,59]. For instance, crosstalk between integrin and tumor growth factor (TGF)-β pathways has been found to be related to several human pathologies, including systemic sclerosis, idiopathic pulmonary fibrosis, chronic obstructive pulmonary disease, and cancer [60]. Finally, disease modules from different disorders might also overlap, providing hypotheses to explain comorbidities. Modularity of diseases can be exploited for the development of more efficient therapeutic strategies [61,62] and to improve disease classification [63].

Networks integrate genetic and environmental factors

Several lines of evidence suggest that gene–environment interactions play significant roles in diseases such as asthma [64], cancer, unipolar depressive disorders, ischemic heart disease, and cerebrovascular disease, among many others [65]. The environmental factors that modulate diseases include allergens, air pollution, cigarette smoke, and viruses, as well as therapeutic drugs that produce adverse effects [66]. Gene–environment interactions can be modeled within the framework of network biology. To test the hypothesis that proteins targeted by environmental factors are in the network vicinity of disease-associated proteins, network distance measures, such as the average shortest path (Box 1) or pathway-discovery algorithms [67–69], can be applied to networks (Figure 1d). As an example of this approach, the analysis of the risk of cardiac arrhythmias of five antipsychotic drugs indicated that the average shortest path length between pairs of drug- and disease-associated proteins was significantly smaller than would be expected by chance [66]. This suggests that proteins in the network vicinity of drug targets are more likely to be involved in adverse effects and provides a hypothesis to explain cardiac arrhythmias caused by antipsychotic drugs. A similar approach looking at associations between a group of genetic diseases and viral infections [70] found that viral targets in the host interactome were in the local neighborhood of disease proteins, and a study focusing on the relation between DNA tumor virus and cancer showed rewiring of the host network in response to viral perturbations [71]. Together, these studies exemplify the application of networks to both integrate and analyze genetic and environmental contributions to human diseases.

Interpretation of GWAS and whole-genome sequencing data

GWAS have accelerated the discovery of genes associated with complex diseases, but the translation of these findings to an understanding of disease pathogenesis has proven difficult due to epistasis [72], among other issues [73]. From a systems-level perspective, epistasis can be viewed as the consequence of the functional effect of gene variants on the entire network of interactions in which they are immersed in the cell [74]. Thus, biological epistasis can be explained by the interaction of proteins in the context of biological networks and pathways [75]. Network analysis offers a powerful set of methods to study this genetic phenomenon at a systems-wide level. Several authors proposed the use of PINs to analyze epistasis in relation to human disease, the rationale being that knowledge of biological networks can be used to narrow the search for epistasis between loci. In addition, this kind of approach offers the possibility of a biological interpretation of the interaction between genes, contrasting with purely statistical models [75–79]. Although these novel, network-based methodologies are proving useful for the identification of interactions between individual markers, they are limited to SNPs that have a clear effect on protein function, omitting most of the variants located in noncoding regions. To overcome this limitation, computational methods predicting the effects of SNPs in gene regulatory regions could be used to include all the variants found in the GWAS studies [8].

More recently, NGS has opened up the door to sequencing the genome of an individual, raising not only several regulatory, methodological, technological, and educational issues, but also fundamental questions, such as which variants are clinically actionable, which ones should be the focus of genomic diagnoses, and what kind of information should be reported back to the patient [80]. The interpretation of personal genomic data is one of the great challenges ahead in personalized medicine. Several studies have highlighted the difficulty in identifying clinically relevant variants from whole-genome sequence data using the information currently available in databases, such as

Box 2. Can network analysis aid in the interpretation of personal genomes?

A recent study applied whole-genome sequencing to characterize a rare tumor with the goal of guiding personalized therapeutic intervention in the absence of an established treatment [107]. The integration of copy number, expression, and mutational data with pathway analysis suggested that tumor growth was driven by activation of the RET oncogene. The patient was then treated with sunitinib, a drug that inhibits RET but is not indicated for this tumor type, leading to a stabilization of the disease for 4 months. Recurrence of tumor growth prompted new sequencing analysis, yielding new genetic alterations that explained drug resistance through activation of the mitogen-activated protein kinase (MAPK) and AKT pathways.

Whole-genome sequencing is also being used to assess disease risk in healthy individuals. A pioneering study combining genomic, transcriptomic, proteomic, and autoantibody profiles was performed on a healthy individual to monitor dynamical changes in molecular and medical phenotypes over a 14-month period [4]. Network and pathway analysis helped identify distinct biological processes characteristic of two viral infections that occurred during the length of the study, as well as of the development of type 2 diabetes mellitus. Although network analysis of PINs is not yet an integral part of the interpretation of personal genomic data, these studies suggest that its application would aid in the identification of etiological modules perturbed in the diseases and could guide strategies for therapeutic interventions.

Online Mendelian Inheritance in Man (OMIM), or the literature [6,80,81]. This is challenging not only for mutations identified in Mendelian disorders, but also particularly in studies focusing in healthy individuals where putative harmful sequence variants are being discovered [4,6]. Some early studies suggest that network analysis can aid in the interpretation of genomic data (Box 2).

Concluding remarks

Network biology has emerged as an integrative and systems-level approach to aid in the interpretation of genome data in the context of health and disease. However, some challenges remain to realizing the full potential of network biology for understanding human diseases (Box 4). One of the current limitations of network biology is the coverage and quality of interactome data. The data incompleteness of the human PIN poses limitations to any study of network properties of disease genes. In addition, the availability of condition-specific interactomes that are more representative of the interactions of the proteins in a given tissue or under certain conditions will improve the significance of such analysis (Box 3).

The role that disease proteins play in the PIN remains a matter of debate. It is interesting to note that most of the first studies proposing disease genes as hubs were performed on cancer [27,34,35]. Although many cancer types are inherited, the cell transformation process that occurs during cancer development results from the accumulation of somatic mutations in different genes. Thus, it would be interesting to evaluate the network properties of disease genes taking into account the type of mutation (somatic or germ-line) of disease genes. This type of experiment may offer more definitive evidence as to the role of disease proteins in PINs. In addition, differential network mapping should offer more insight into the network rewiring that occurs during disease. This approach would also enable monitoring changes in the role of the individual nodes (hubs, bottlenecks, etc.) and changes in the global

Box 3. Condition-specific interactomes

Most of the studies on the involvement of protein interactions in human disease to date have been performed using aggregate PINs. Aggregate PINs are conglomerates of interactions obtained by integrating different data sets. They represent the set of possible physical interactions between the tested proteins but, due to the experimental methods used to identify the interactions, the aggregate interactomes lack spatiotemporal resolution [108]. Are the network properties of disease genes measured in aggregate PINs maintained in condition-specific interactomes? Although this has not been fully addressed yet, some studies show differences in size and topology between aggregate and condition-specific interactomes [41,108]. Condition-specific interactomes can be obtained by integrating protein interaction data with gene or protein expression information [41,108,109] (Figure I). In this kind of approach, two proteins are assumed to interact in a given condition only if both proteins (or their corresponding mRNA) are expressed in this condition. By using this methodology, one group obtained 86 organ- and cell type-specific PINs, which comprised between 1% and 25% of the interactions from the original databases [41]. In a study of the molecular pathways associated with the proteins targeted by HIV and hepatitis C (HCV) viruses, meaningful results were obtained when using tissue-specific subnetworks, but not with the aggregate PINs. This last point is particularly important to the study of human disease, where it may be necessary to reassess the network topology properties of proteins encoded by disease genes in interactomes with spatiotemporal resolution in both healthy and disease states. This will provide a more realistic scenario and might help to resolve the discrepancies found in previous studies [15,37–39].

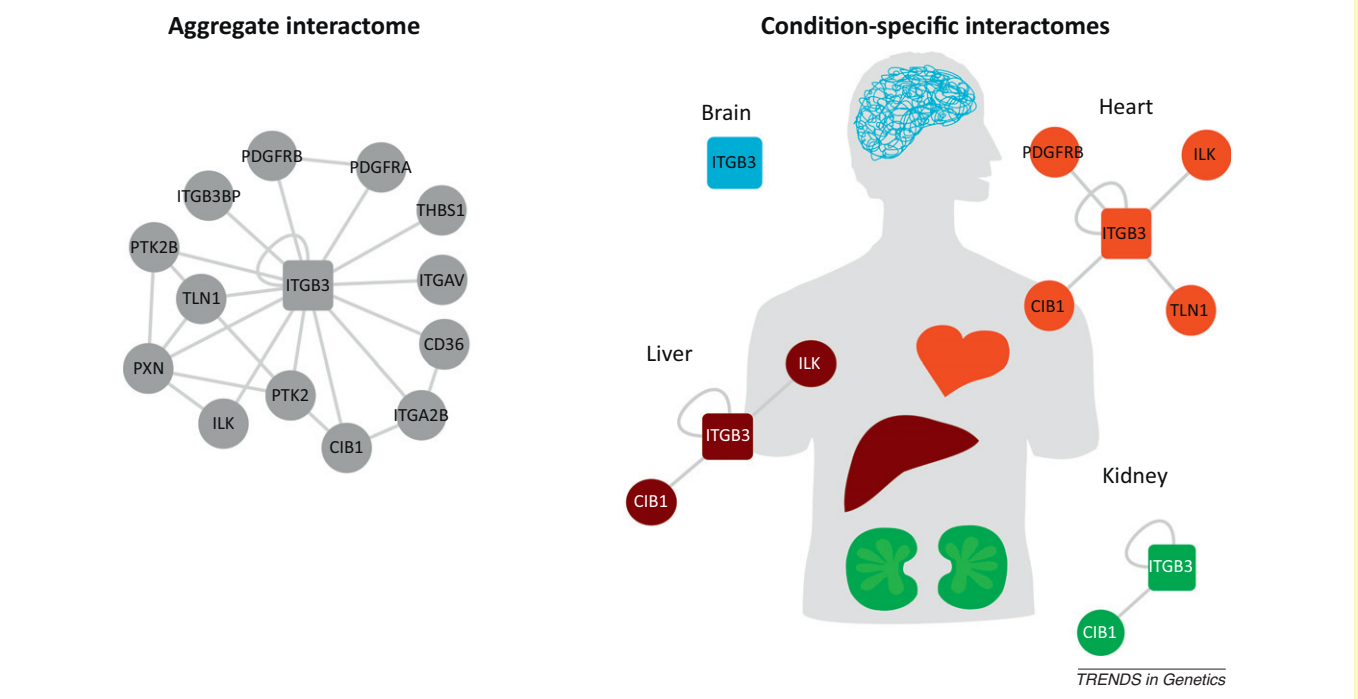

Figure I. Aggregate and condition-specific interactomes. The protein interaction network (PIN) around integrin beta-3 (ITGB3) is used as an example, as obtained from the Human Integrated Protein–Protein Interaction Reference (HIPPIE) [89]. In the aggregate interactome, ITGB3 interacts with 13 proteins, whereas in the PINs obtained by integration of gene expression data and protein interactions for brain, heart, kidneys, and liver, only a subset of the interactions are maintained.

Box 4. Outstanding questions

- Can network biology help in the identification of clinically actionable research findings (e.g., biomarkers and drug targets)?
- Can network biology explain penetrance and expressivity in genetic diseases?
- What can network analysis tell us about the role of gene epistasis in disease?
- Are network-perturbation models able to explain underlying mechanisms of Mendelian and complex disorders?
- How does the location of a protein within a network influence the phenotypic consequences of a mutation in that protein? Can these findings be generalized for proteins in different locations (central and peripheries)? Can these findings be generalized for all types of mutation (germ-line or somatic)?
- Can we understand disease causation by the combined effect of environmental and genetic perturbations using network biology?

topology of the PIN, as well as how all these alterations correlate with cell function. By fully characterizing network rewiring in disease, a deeper understanding of how sequence variation shapes cellular networks and leads to observed phenotypic changes would be gained.

One caveat to these methods is that they are costly and time consuming, limiting their practical utility for the interpretation of personal genome data. Here, bioinformatic methods for the prediction of the impact of sequence variation could play an important role. Although there are several methods and tools already available for that purpose [8], they need to be improved to cope with the current demands of GWAS and personal genome data analysis. The prediction of the functional effect of sequence variants and mutations is an open research area, as evidenced by recent conferences on the topic ([82]; http://www.unbsj.ca/sase/csas/data/aimm2012/index.html). Existing methods focus mainly on the prediction of the functional consequences of mutations at the protein level [8]. More importantly, with the exception of some first attempts [83], there is currently no approach able to evaluate the consequences of sequence variations at the network level. Analogous to the differential network mapping methodologies discussed above, there is room for the development of *in silico* methods to assess network rewiring as a consequence of disease-associated sequence variation. Such methodologies could predict, for instance, if a mutation leads to a node removal or an edgetic perturbation, how the structure of the network is modified in the presence of such perturbations, and how functional modules are modified or reconnected due to disease-associated mutations. The availability of such methodology would aid in the interpretation of personal genome data by providing mechanistic hypothesis of the effect of sequence variation and identify potential targets for personalized therapeutic intervention.

Acknowledgments

The author thanks the editor and reviewers for valuable criticisms on the manuscript. This work has received support from the IMI Joint Undertaking under grant agreement no. 115002, eTOX, resources of which comprise financial contribution from the EU FP7 (FP7/2007-2013) and European Federation of Pharmaceutical Industries and Associations (EFPIA) companies' in kind contribution; the IMI Joint Undertaking under grant agreement no. 115191, OpenPhacts, resources of which comprise financial contribution from the EU FP7 (FP7/2007-2013) and EFPIA companies' in kind contribution; and Instituto de Salud Carlos III FEDER [CP10/00524]. The Research Programme on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB).

References

1 Ku, C.S. *et al.* (2010) The pursuit of genome-wide association studies: where are we now? *J. Hum. Genet.* 55, 195–206
2 Manolio, T. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature* 461, 747–753
3 Ku, C.S. *et al.* (2011) Revisiting Mendelian disorders through exome sequencing. *Hum. Genet.* 129, 351–370
4 Chen, R. *et al.* (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148, 1293–1307
5 Ormond, K.E. *et al.* (2010) Challenges in the clinical application of whole-genome sequencing. *Lancet* 375, 1749–1751
6 Ball, M.P. *et al.* (2012) A public resource facilitating clinical use of genomes. *Proc. Natl. Acad. Sci. U.S.A.* 109, 11920–11927
7 Wang, J. *et al.* (2012) Identification of aberrant pathways and network activities from high-throughput data. *Brief. Bioinform.* 13, 406–419
8 Cooper, G.M. and Shendure, J. (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* 12, 628–640
9 Ober, C. and Vercelli, D. (2011) Gene–environment interactions in human disease: nuisance or opportunity? *Trends Genet.* 27, 107–115
10 Barabasi, A. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* 12, 56–68
11 Johnson, A.D. *et al.* (1981) lambda Repressor and cro- components of an efficient molecular switch. *Nature* 294, 217–223
12 Sardiu, M.E. and Washburn, M.P. (2011) Building protein–protein interaction networks with proteomics and informatics tools. *J. Biol. Chem.* 286, 23645–23651
13 Stumpf, M.P.H. *et al.* (2008) Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. U.S.A.* 105, 6959–6964
14 Venkatesan, K. *et al.* (2009) An empirical framework for binary interactome mapping. *Nat. Methods* 6, 83–90
15 Goh, K. *et al.* (2007) The human disease network. *Proc. Natl. Acad. Sci. U.S.A.* 104, 8685
16 Feldman, I. *et al.* (2008) Network properties of genes harboring inherited disease mutations. *Proc. Natl. Acad. Sci. U.S.A.* 105, 4323–4328
17 Bauer-Mehren, A. *et al.* (2011) Gene–disease network analysis reveals functional modules in Mendelian, complex and environmental diseases. *PLoS ONE* 6, e20284
18 Chen, Y. *et al.* (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature* 452, 429–435
19 Yang, X. *et al.* (2009) Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. *Nat. Genet.* 41, 415–423
20 Jones, S. *et al.* (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* 321, 1801–1806
21 Emilsson, V. (2008) Genetics of gene expression and its effect on disease. *Nature* 452, 423–428
22 Lee, D.S. *et al.* (2008) The implications of human metabolic network topology for disease comorbidity. *Proc. Natl. Acad. Sci. U.S.A.* 105, 9880–9885
23 Park, J. *et al.* (2009) The impact of cellular networks on disease comorbidity. *Mol. Syst. Biol.* 5, 262
24 Park, S. *et al.* (2011) Protein localization as a principal feature of the etiology and comorbidity of genetic diseases. *Mol. Syst. Biol.* 7, 494
25 Yildirim, M.A. *et al.* (2007) Drug–target network. *Nat. Biotechnol.* 25, 1119–1126
26 Yamanishi, Y. *et al.* (2008) Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24, i232–i240
27 Taylor, I.W. *et al.* (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* 27, 199–204
28 Marbach, D. *et al.* (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804
29 Kholodenko, B. *et al.* (2012) Computational approaches for analyzing information flow in biological networks. *Sci. Signal.* 5, 1
30 Wang, X. *et al.* (2011) Network-based methods for human disease gene prediction. *Brief. Funct. Genomics* 10, 280–293
31 Jeong, H. *et al.* (2001) Lethality and centrality in protein networks. *Nature* 411, 41–42

32 Han, J.J. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* 430, 88–93
33 Zhu, X. *et al.* (2007) Getting connected: analysis and principles of biological networks. *Genes Dev.* 21, 1010–1024
34 Wachi, S. *et al.* (2005) Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* 21, 4205–4208
35 Jonsson, P.F. and Bates, P.A. (2006) Global topological features of cancer proteins in the human interactome. *Bioinformatics* 22, 2291–2297
36 Xu, J. and Li, Y. (2006) Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics* 22, 2800–2805
37 Jin, W. *et al.* (2012) A systematic characterization of genes underlying both complex and Mendelian diseases. *Hum. Mol. Genet.* 21, 1611–1624
38 Podder, S. and Ghosh, T.C. (2010) Exploring the differences in evolutionary rates between monogenic and polygenic disease genes in human. *Mol. Biol. Evol.* 27, 934–941
39 Cai, J.J. *et al.* (2010) Broker genes in human disease. *Genome Biol. Evol.* 2, 815–825
40 Agarwal, S. *et al.* (2010) Revisiting date and party hubs: novel approaches to role assignment in protein interaction networks. *PLoS Comput. Biol.* 6, 1–12
41 Lopes, T.J.S. *et al.* (2011) Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics* 27, 2414–2421
42 Yu, H. *et al.* (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.* 3, e59
43 Zhong, Q. *et al.* (2009) Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.* 5, 321
44 Strachan, T. and Read, A.P. (1999) Molecular pathology. In *Human Molecular Genetics (2nd edn)* (Strachan, T. and Read, A.P., eds), Wiley-Liss available from: (http://www.ncbi.nlm.nih.gov/books/NBK7574/)
45 Wang, X. *et al.* (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* 30, 159–164
46 Ideker, T. and Krogan, N.J. (2012) Differential network biology. *Mol. Syst. Biol.* 8, 1
47 Zaghloul, N.A. and Katsanis, N. (2010) Functional modules, mutational load and human genetic disease. *Trends Genet.* 26, 168–176
48 Zhao, S. and Li, S. (2010) Network-based relating pharmacological and genomic spaces for drug target identification. *PLoS ONE* 5, 11764
49 Azmi, A. *et al.* (2010) Proof of concept: a review on how network and systems biology approaches aid in the discovery of potent anticancer drug combinations. *Mol. Cancer Ther.* 9, 3137–3144
50 Navlakha, S. and Kingsford, C. (2010) The power of protein interaction networks for associating genes with diseases. *Bioinformatics* 26, 1057–1063
51 Oti, M. and Brunner, H.G. (2007) The modular nature of genetic diseases. *Clin. Genet.* 71, 1–11
52 Lim, J. (2006) A protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* 125, 801–814
53 D'Andrea, A.D. and Grompe, M. (2003) The Fanconi anaemia/BRCA pathway. *Nat. Rev. Cancer* 3, 23–34
54 Van Reeuwijk, J. *et al.* (2005) Glyc-O-genetics of Walker–Warburg syndrome. *Clin. Genet.* 67, 281–289
55 Lage, K. *et al.* (2007) A human phenome–interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* 25, 309–316
56 Cerami, E. *et al.* (2010) Automated network analysis identifies core pathways in glioblastoma. *PLoS ONE* 5, e8918
57 Walsh, T. and King, M. (2007) Ten genes for inherited breast cancer. *Cancer Cell* 11, 103–105
58 Shao, L. *et al.* (2012) Dynamic network of transcription and pathway crosstalk to reveal molecular mechanism of MGd-treated human lung cancer cells. *PLoS ONE* 7, 31984
59 Xu, Y. *et al.* (2010) Prediction of human protein–protein interaction by a mixed Bayesian model and its application to exploring underlying cancer-related pathway crosstalk. *J. R. Soc. Interface* 8, 555–567
60 Margadant, C. and Sonnenberg, A. (2010) Integrin-TGF-beta crosstalk in fibrosis, cancer and wound healing. *EMBO Rep.* 11, 97–105
61 Fernández, A. and Sessel, S. (2009) Selective antagonism of anticancer drugs for side-effect removal. *Trends Pharmacol. Sci.* 30, 403–410
62 Berger, S.I. and Iyengar, R. (2009) Network analyses in systems pharmacology. *Bioinformatics* 25, 2466–2472
63 Loscalzo, J. *et al.* (2007) Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Mol. Syst. Biol.* 3, 124
64 von Mutius, E. (2009) Gene–environment interactions in asthma. *J. Allergy Clin. Immunol.* 123, 3–11
65 Hunter, D.J. (2005) Gene–environment interactions in human diseases. *Nat. Rev. Genet.* 6, 287–298
66 Bauer-Mehren, A. *et al.* (2012) Automatic filtering and substantiation of drug safety signals. *PLoS Comput. Biol.* 8, 1002457
67 Tuncbag, N. *et al.* (2012) Simultaneous reconstruction of multiple signaling pathways via the prize-collecting Steiner Forest problem. *Lect. Notes Comput. Sci.* 7262, 287–301
68 Gitter, A. *et al.* (2011) Discovering pathways by orienting edges in protein interaction networks. *Nucleic Acids Res.* 39, e22
69 Supper, J. *et al.* (2009) BowTieBuilder: modeling signal transduction pathways. *BMC Syst. Biol.* 3, 67
70 Gulbahce, N. *et al.* (2012) Viral perturbations of host networks reflect disease etiology. *PLoS Comput. Biol.* 8, 1002531
71 Rozenblatt-Rosen, O. *et al.* (2012) Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. *Nature* 487, 491–495
72 Carlborg, O. and Haley, C.S. (2004) Epistasis: too often neglected in complex trait studies? *Nat. Rev. Genet.* 5, 618–625
73 Mitchell, K.J. (2012) What is complex about complex disorders? *Genome Biol.* 13, 237
74 Tyler, A.L. *et al.* (2009) Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *Bioessays* 31, 220–227
75 Pattin, K. and Moore, J. (2008) Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases. *Hum. Genet.* 124, 19–29
76 Akula, N. *et al.* (2011) A network-based approach to prioritize results from genome-wide association studies. *PLoS ONE* 6, 24220
77 Jia, P. *et al.* (2012) Network-assisted investigation of combined causal signals from genome-wide association studies in schizophrenia. *PLoS Comput. Biol.* 8, 1002587
78 Bakir-Gungor, B. and Sezerman, O.U. (2011) A new methodology to associate SNPs with human diseases according to their pathway related context. *PLoS ONE* 6, 26277
79 Emily, M. *et al.* (2009) Using biological networks to search for interacting loci in genome-wide association studies. *Eur. J. Hum. Genet.* 17, 1231–1240
80 Biesecker, L.G. *et al.* (2012) Next-generation sequencing in the clinic: are we ready? *Nat. Rev. Genet.* 13, 818–824
81 Berg, J.S. *et al.* (2012) An informatics approach to analyzing the incidentalome. *Genet. Med.* http://dx.doi.org/10.1038/gim.2012.112
82 Bromberg, Y. and Capriotti, E. (2012) SNP-SIG Meeting 2011: identification and annotation of SNPs in the context of structure, function, and disease. *BMC Genomics* 13, 1–2
83 Bauer-Mehren, A. *et al.* (2009) From SNPs to pathways: integration of functional effect of sequence variations on models of cell signalling pathways. *BMC Bioinformatics* 10, S6
84 Griffiths, A.J.F. *et al.* (1999) Penetrance and expressivity. In *Modern Genetic Analysis* (Griffiths, A.J.F. *et al.*, eds), W.H. Freeman available from: (http://www.ncbi.nlm.nih.gov/books/NBK21352/)
85 Strachan, T. and Read, A.P. (1999) Genes in pedigrees. In *Human Molecular Genetics (2nd edn)* (Strachan, T. and Read, A.P., eds), Wiley-Liss available from: (http://www.ncbi.nlm.nih.gov/books/NBK7573/)
86 Cordell, H. (2009) Detecting gene–gene interactions that underlie human diseases. *Nat. Rev. Genet.* 10, 392–404
87 Handel, A.E. *et al.* (2010) Epigenetics: molecular mechanisms and implications for disease. *Trends Mol. Med.* 16, 7–16
88 Bader, G.D. *et al.* (2006) Pathguide: a pathway resource list. *Nucleic Acids Res.* 34, D504–D506
89 Schaefer, M.H. *et al.* (2012) HIPPIE: integrating protein interaction networks with experiment based quality scores. *PLoS ONE* 7, e31826

90 Stark, C. *et al.* (2011) The BioGRID interaction database: 2011 update. *Nucleic Acids Res.* 39, D698–D704
91 Kerrien, S. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* 40, D841–D846
92 Szklarczyk, D. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* 39, D561–D568
93 Ceol, A. *et al.* (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.* 38, D532–D539
94 Prasad, T.S.K. *et al.* (2009) Human protein reference database: 2009 update. *Nucleic Acids Res.* 37, D767–D772
95 Hu, Z. *et al.* (2009) VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Res.* 37, W115–W121
96 Smoot, M.E. *et al.* (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431–432
97 Kitano, H. *et al.* (2005) Using process diagrams for the graphical representation of biological networks. *Nat. Biotechnol.* 23, 961–966
98 Bastian, M. *et al.* (2009) Gephi: an open source software for exploring and manipulating networks, In *International AAAI Conference on Weblogs and Social Media*, pp. 361–362
99 Ellson, J. *et al.* (2001) Graphviz: open Source Graph Drawing Tools. *Lect. Notes Comput. Sci.* 2265, 483–484
100 Köhler, J. *et al.* (2006) Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics* 22, 1383–1390
101 Breitkreutz, B.J. *et al.* (2003) Osprey: a network visualization system. *Genome Biol.* 4, R22
102 Longabaugh, W.J.R. *et al.* (2009) Visualization, documentation, analysis, and communication of large scale gene regulatory networks. *Biochim. Biophys. Acta* 1789, 363
103 Dogrusoz, U. *et al.* (2006) PATIKAweb: a Web interface for analyzing biological pathways through advanced querying and visualization. *Bioinformatics* 22, 374–375
104 Theocharidis, A. *et al.* (2009) Network visualization and analysis of gene expression data using BioLayout Express3D. *Nat. Protoc.* 4, 1535–1550
105 Pavlopoulos, G.A. *et al.* (2008) Arena3D: visualization of biological networks in 3D. *BMC Syst. Biol.* 2, 104
106 Kozhenkov, S. *et al.* (2010) BiologicalNetworks 2.0: an integrative view of genome biology data. *BMC Bioinformatics* 11, 610
107 Jones, S. *et al.* (2010) Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. *Genome Biol.* 11, R82
108 Souiai, O. *et al.* (2011) Functional integrative levels in the human interactome recapitulate organ organization. *PLoS ONE* 6, 22051
109 Bossi, A. and Lehner, B. (2009) Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.* 5, 1