BRIEFINGS IN FUNCTIONAL GENOMICS. VOL 10. NO 5. 280–293 doi:10.1093/bfgp/elr024

Network-based methods for human disease gene prediction

Xiujuan Wang, Natali Gulbahce and Haiyuan Yu

Advance Access publication date 15 July 2011

Abstract

Despite the considerable progress in disease gene discovery, we are far from uncovering the underlying cellular mechanisms of diseases since complex traits, even many Mendelian diseases, cannot be explained by simple genotype–phenotype relationships. More recently, an increasingly accepted view is that human diseases result from perturbations of cellular systems, especially molecular networks. Genes associated with the same or similar diseases commonly reside in the same neighborhood of molecular networks. Such observations have built the basis for a large collection of computational approaches to find previously unknown genes associated with certain diseases. The majority of the methods are based on protein interactome networks, with integration of other large-scale genomic data or disease phenotype information, to infer how likely it is that a gene is associated with a disease. Here, we review recent, state of the art, network-based methods used for prioritizing disease genes as well as unraveling the molecular basis of human diseases.

Keywords: *human diseases; disease network; disease gene prediction; protein–protein interaction; molecular network*

INTRODUCTION

Many ground-breaking discoveries of genes associated with human diseases and their molecular bases have dramatically increased our understanding of the development of diseases over the last decades [1]. Uncovering the underlying molecular basis of diseases has become incredibly valuable in the prevention, diagnosis and treatment of diseases. Despite the steady increase in discovering disease-associated genes, there is still a large fraction of diseases without a known molecular basis. Currently, there are over 1700 diseases with no known molecular basis curated in the OMIM (Online Mendelian Inheritance in Man) database as this review is being written. Even for those diseases for which there is a partial knowledge of a molecular basis, a large proportion of their associated genes are still not known. It has been reported that the genes established to be associated with diseases such as cancer and type 2 diabetes only represent a very small proportion of the incidences [2, 3]. Hence, the majority of disease genes still remain underneath the tip of the iceberg.

Many approaches have been dedicated to the discovery of candidate genes [4]. Traditional genetic mapping methods include linkage analysis and genome-wide association studies (GWAS) of Mendelian diseases and complex traits. While GWAS are powerful and fruitful, they face challenges in narrowing down the long lists of candidate genes [5]. Furthermore, human diseases generally

Corresponding author. Haiyuan Yu, Department of Biological Statistics and Computational Biology and Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY 14850, USA. Tel: 607 255 0259; Fax: 607 255 5961; E-mail: haiyuan.yu@cornell.edu

Xiujuan Wang is a postdoctoral fellow in the Department of Molecular Biology and Genetics and Weill Institute for Cell and Molecular Biology at Cornell University. Her research focuses on discovery of unknown human disease genes through decoding interactome networks.

Natali Gulbahce is a research fellow at the Department of Cellular and Molecular Pharmacology, University of California, San Francisco, USA. Her research focus is network-based disease pathway identification and the impact of perturbations on cellular networks.

Haiyuan Yu is an assistant professor in the Department of Biological Statistics and Computational Biology and Weill Institute for Cell and Molecular Biology at Cornell University. His laboratory aims to understand gene functions and their relationships within complex molecular networks and how perturbations to such systems may lead to various human diseases using both high-throughput experimental and integrative computational methodologies.

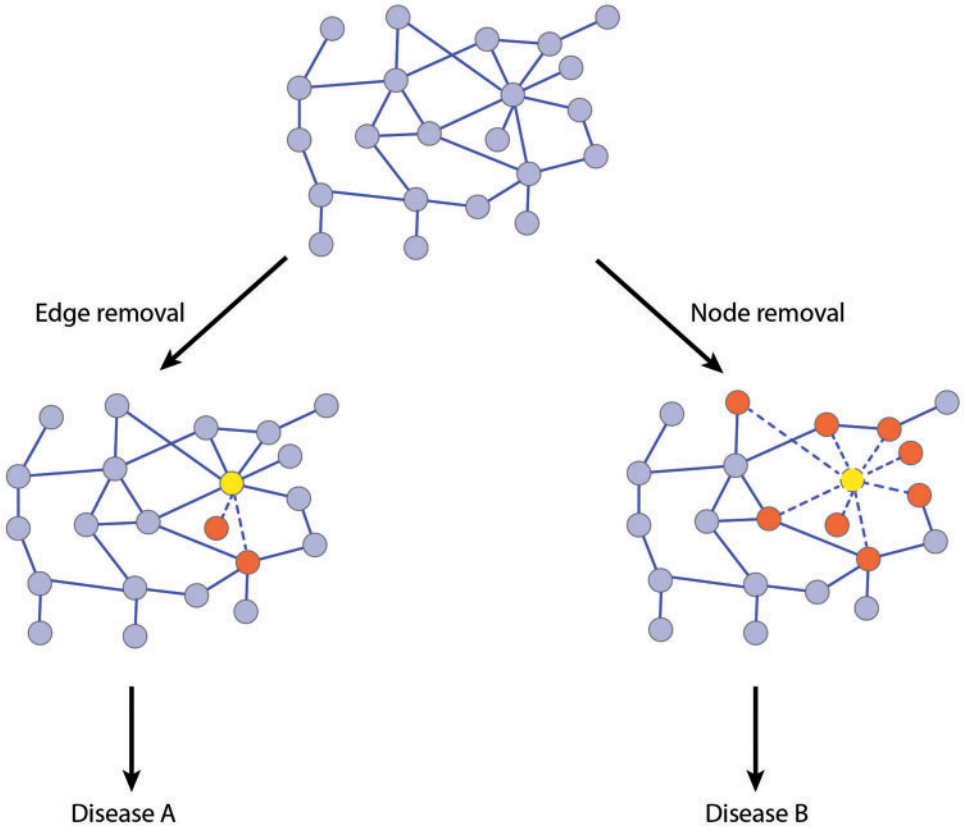

Figure 1: Perturbations in molecular networks disrupt biological pathways and result in human diseases. Mutations in a node (highlighted in yellow) cause different types of perturbations in molecular networks with directly affected neighbors shown in orange. Disease A is triggered as the result of an edge removal, and disease B is developed due to the node removal. The two diseases are not necessarily the same, but may share similarity in phenotypes.

do not follow the simple genotype–phenotype relationship hypothesis, but are rather the consequences of perturbations in the molecular networks induced by various factors such as genetic mutations, epigenetic changes and pathogens [6]. The efforts in unraveling the properties of disease genes in molecular networks have shown that genes associated with the same or similar diseases, tend to reside in the same neighborhood in these networks and form physical and/or functional modules [7–9]. These findings became the basis for the development of computational approaches for predicting and prioritizing candidate disease genes. In this review, we focus on state of the art approaches in this rapidly growing field that are built on interactome and protein–protein interaction (PPI) networks in particular.

MOLECULAR NETWORKS

Molecular networks, including PPI, metabolic, regulatory, genetic and co-expression networks, have been steadily constructed experimentally to characterize the physical and/or functional interactions between biomolecules (see [10, 11] for comprehensive reviews of these networks). Perturbations in these wiring diagrams may trigger particular phenotypes in both monogenic and polygenic diseases, including tumors (Figure 1). Deciphering the properties of these networks will offer a much deeper understanding into complex genotype–phenotype relationships. Molecular networks can be subdivided into two categories: interactome networks (metabolic, PPI and gene regulatory networks) that represent physical or biochemical interactions between macromolecules, and functional networks (transcription profiling, phenotypic profiling and genetic interaction networks) that display functional relationships or similarities between genes and gene products [11]. These networks are commonly displayed as a graph with nodes as molecules and directed or undirected edges as links between them [12] (Box 1 for basic graphic concepts of networks). PPI networks usually have undirected edges, representing the physical interactions between the proteins (i.e. nodes). On

Box1 Graphic concepts of the molecular network

A molecular network is usually depicted as a set of nodes connected by directed (regulatory and metabolic networks) or undirected (PPI networks) edges.

Path: a sequence of nodes starting from one node to another node connected by edges;
Length of a path: the number of edges in the path;
Shortest path: a path between two nodes with smallest length;
Degree: the number of edges a node is connected with;
Kth-order neighbor: the nodes whose distance to a node is k;
Betweenness: the number of shortest paths that run through a node;
Hubs: a node with high degree;
Party hubs: highly co-expressed with the connected nodes;
Date hubs: interact with the connected nodes at different time and/or locations;

the contrary, gene regulatory networks are constructed with the nodes connected by directed edges representing physical binding of one node (transcription factor) to the other nodes (DNA regulatory elements). In metabolic networks, the nodes are biochemical metabolites and the edges, either directed or undirected, represent reactions or enzymes catalyzing the reactions to convert one node into another. Functional networks of transcription profiling, phenotypic profiling and genetic interaction, all have nodes representing genes, but edges representing highly correlated co-expression, highly correlated phenotypic profiles and known genetic interactions, respectively. The current status as well as approaches for constructing these networks can be found in two recent comprehensive reviews [10, 11].

Many proteins carry out their functions through interacting with other proteins. Two main high-throughput technologies have been advanced and are successful in producing a large number of PPIs in humans: (i) a high-throughput yeast two-hybrid (Y2H) system has been developed to systematically screen for direct binary interactions between protein pairs [13–15] and (ii) high-throughput affinity purification (AP) followed by mass spectrometry (MS) approaches have been employed to identify protein complexes in humans [16, 17]. Significant efforts have also been made to search through the literature and curate the interactions that have been reported by small-scale experiments, as was done in a large number of databases such as Human Protein Reference Database (HPRD), Molecular Interaction database (MINT), Biological General Repository for Interaction data sets (BioGRID), Biomolecular Interaction Network Database (BIND) and IntAct [18–22]. Despite the fact that errors and biases are still present in this incomplete human PPI network [23], the nonstop exertion in constructing high-coverage and high-quality PPI networks has made the computational prediction of disease genes possible. It should be noted that, although in this review, we focus specifically on decoding PPI networks for discovery of disease genes, analogous principles can be applied to the other types of networks mentioned above.

THE PROPERTIES OF DISEASE GENES IN PPI NETWORKS

Most molecular networks are scale-free such that the distribution of node connectivity (number of neighbors) follows a power law rather than a Poisson distribution. In such scale-free networks, the majority of nodes have few links while other nodes, so called hubs, have a much higher degree of linkages. In model organisms, hub proteins have been reported as essential and more abundant, and they generally display a greater diversity of phenotypes in knockouts when compared to nonhub proteins [24–28]. These findings lead to the question of whether or not disease-associated genes in humans tend to encode hubs in cellular networks. The analysis of differentially expressed genes in cancer suggested that up-regulated genes in lung squamous cancer tissues have significantly higher connectivity in the PPI network [29]. A similar conclusion was drawn by Jonsson and Bates [30] that cancer-related proteins have about twice the interaction partners when compared with proteins unrelated to cancer. However, these observations may be the result of a bias, in that cancer proteins are often much better studied. Goh *et al.* [8] showed that disease gene products displayed more of a tendency to encode hubs in the PPI network than nondisease gene products. However, further investigation demonstrated that only essential disease genes were associated with

hubs and were widely expressed, while nonessential disease genes did not demonstrate these characteristics [8, 9]. Another observation is that network neighbors of disease genes tend to be involved in the same or similar diseases. Genes causing similar disease phenotypes are often functionally related and form a biological module such as a protein complex or pathway [7]. Goh *et al.* [8] showed that genes associated with the same disorder have significantly higher gene ontology (GO) homogeneity than random expectation as well as an increased tendency to be co-expressed. It has been shown that genes causing the same phenotype tend to form topological clusters [9]. These distinct features of disease genes as revealed by interactome and functional networks can be adopted to identify functionally similar genes in addition to uncharacterized disease genes.

DISEASE GENE PREDICTION

Proximity of proteins in the PPI network

Current approaches for disease gene prioritization mostly rely on the proximity of candidate genes to known disease genes within interactome networks using different scoring strategies. The underlying assumption is 'guilt-by-association', in that, genes that are physically or functionally close to each other tend to be involved in the same biological pathways and have similar effects on phenotypes [31, 32]. Hence, a key step is to measure the distance between candidate genes and known disease genes in the PPI network, for which an increasing number of approaches have been developed [33, 34]. Here, we focus on three main categories: local distance measurements, global distance measurements and other graphic clustering methods to measure pair-wise protein closeness in a network for prioritizing candidate genes (Table 1).

The most straightforward approach is to assess whether two proteins are connected directly in a network, so called direct neighbor counting. The count for any protein pair is 1 if the two proteins are directly connected by an edge, with count of 0 otherwise (see Table 1). The more disease genes that a candidate gene is directly connected to, the more probable it is that the candidate is associated with the same disease. Oti *et al.* [35] predicted disease-causing genes in known disease loci by counting the number of known causative genes in their direct network neighbors (Table 2). They achieved an

Table 1: Approaches for measuring proximity of elements in PPI networks

Method	Function	Description	References
Direct neighborhood	$N_{uv} = \begin{cases} 1, & \text{if } \exists\, E_{uv} \\ 0, & \text{otherwise} \end{cases}$	The count N_{uv} for protein pair u and v is 1 if they are directly connected by an edge E_{uv}, and is 0 otherwise.	[35, 38, 39, 43, 45, 50, 51, 55]
Shortest path length	$D_{uv} = L_{uv}, \quad \text{where } L_{uv} \leq L'_{uv}$	The distance D_{uv} between protein u and v is the shortest path length L_{uv}. L'_{uv} is the length of any possible path connecting protein u and v.	[37, 44, 55, 77]
Diffusion kernel	$K = e^{-\beta L}$	The diffusion kernel K of the graph is the function of Laplacian L, the difference of the degree matrix and the adjacency matrix, with parameter β as the control of diffusion magnitude.	[38]
Random walk with restart	$P^t = (1 - r)WP^{t-1} + rP^0$	The random walk with restart is an iterative walker's transition from the current node to a random neighbor with probability r to restart the walk at the source node. W is the adjacency matrix of the graph and P^t is the probability vector being at the nodes at iteration t.	[38, 39, 57]
Propagation flow	$F^t = \alpha W' F^{t-1} + (1 - \alpha)Y$	F^t is the prioritization function representing the relevance of proteins in the network to the seed nodes at iteration t. Each node propagates information received from the previous iteration to its neighbors. Y is the prior information, α is the parameter controling the importance of the prior information and W' is the normalized weight matrix.	[58]

Table 2: Summary of network-based disease gene prediction methods

Class	Reference	Program name	Features	PPI data sources	Proximity measurement of elements in networks	Prediction methods
Approaches solely based on PPI networks	Oti *et al.* [35]	–	PPI	HPRD, human Y2H, fly, worm, yeast	Direct neighbor	Predict a candidate gene as disease gene if it directly interacts with a known disease gene and resides in a known disease locus lacking identified disease genes.
	Kohler *et al.* [38]	–	PPI	HPRD, BIND, BioGrid, IntACT, DIP [78], STRING [79], mapped from worm, mouse, fruitfly and yeast	Direct neighbor, shortest path length, diffusion kernel, random walk with restart	Rank candidate genes based on the proximity scores to known disease genes.
	Navlakha and Kingsford [39]	–	PPI	HPRD, OPHID	Random walk with restart, propagation flow, direct neighbor, graph partitioning, clustering, Markov clustering and their variants	(i) Predict a gene as disease gene if it is located in a locus known to be associated with the disease and the network measurement score is above a threshold; (ii) Combine all 13 closeness measurements for the ensemble of decision trees using a random forest classifier.
Integration of large-scale genomic data and PPI networks	Aerts *et al.* [43]	Endeavour	PPI, TXT, GO, EXP (microarray and EST), PDS, KEGG, TOUCAN, TRANSFAC, SEQ, and others	BIND	Direct neighbor	Rank each candidate gene based on their similarity to known disease genes for each feature, then combine the ranks using order statistics to obtain final rank.
	Franke *et al.* [44]	Prioritizer	PPI, GO, EXP	BIND, HPRD, and large-scale experiments	Shortest path length in the integrated networks	Use a Bayesian classifier to build integrative networks, then score candidate genes based on distance to known disease genes using Gaussian kernel scoring function.
	Radivojac *et al.* [37]	PhenoPred	PPI, GO, Structure, SEQ, DO	HPRD, OPHID	Shortest path length	Employ support vector machines for prediction.
	Linghu *et al.* [45]	–	Curated PPI, Y2H, Masspec, DDI, EXP, PDS, PG, GN, TXT, GO (molecular function and cellular component)	Curated PPI are from HPRD, BIND, BioGrid, IntACT, MIPS [80], DIP, MINT [19], STRING, yeast, worm, fly, mouse-rat	Direct neighbor	Utilize a Bayesian classifier to construct a weighted functional linkage network through integrating large-scale genomic data. The weights of the links to the known disease genes are summed to score candidate genes.
	Karni *et al.* [77]	–	PPI, EXP under disease conditions	HPRD, Y2H	Shortest path length	Find smallest set of genes that cover the disease related genes using the Maximum expectation gene cover algorithm.

(continued)

Table 2 Continued

Class	Reference	Program name	Features	PPI data sources	Proximity measurement of elements in networks	Prediction methods
Integration of disease phenotypic information and PPI network	Lage *et al.* [50]	–	PPI	MINT, BIND, IntAct, Pprel [81], Ecrel [81], Reactome [82]	Direct neighbor (virtual pull down) counting in the evidence-weighted PPI network	Construct candidate complexes by virtual pull down, then score the candidate gene by measuring the similarity between the phenotype caused by the genes in the complex to the disease phenotype.
	Wu *et al.* [55]	CIPHER	PPI	HPRD, OPHID, BIND, MINT	Direct neighbor, shortest path length	Use correlation coefficients as concordance score for each candidate gene based on linear regression of phenotype profile and PPI profile.
	Wu *et al.* [56]	AlignPI	PPI	HPRD	–	Use NetworkBlast algorithm to align the PPI and phenotype networks and obtain high scoring subnetworks (bi-modules). Candidate genes are first assumed to be disease-associated and used for constructing bi-modules. The candidate gene with highest scored bi-module is taken as a positive prediction.
	Care *et al.* [51]	–	PPI, SNPs	BIND, IntAct, BioGrid, MINT	Direct neighbor	Predict deleterious SNPs using random forest, then predict disease genes using the same learning approach based on PPI networks, phenotype similarities and deleterious SNPs.
	Li and Patra [57]	RWRH	PPI	HPRD	Random Walk with Restart on the Heterogeneous network	Use RWRH to score genes and diseases simultaneously by allowing random walker jump between PPI and phenotype networks.
	Vanunu *et al.* [58]	PRINCE	PPI	HPRD and large scale experiments	Network propagation flow in the evidence-weighted PPI network	Use the network propagation method to smooth flow in the PPI network and then use the final converged flow as scores for candidate genes.

PPI, protein–protein interaction; Y2H, yeast two hybrid experiments; PDS, protein domain sharing; PG, phylogenetic profiles; GN, gene neighbor; GO, gene ontology; EXP, gene expression; KEGG, Kyoto Encyclopedia of Genes and Genomes for pathway membership; TOUCAN, *cis*-regulatory modules; TRANSFAC, transcriptional motifs; SEQ, sequence similarity; DO, disease ontology; TXT, literature text mining; Masspec, mass spectrometry; DDI, domain–domain interactions; SNPs, single nucleotide polymorphisms; DIP: Database of Interacting Proteins; STRING: Search Tool for the Retrieval of Interacting Genes/Proteins.

approximately 10-fold enrichment by comparing their candidates to a random selection of candidate genes at the same locus.

Since two proteins can be involved in the same biological pathway without a physical interaction, a number of researchers quantified the closeness of proteins in PPI networks using the shortest path length between them [36, 37]. Krauthammer *et al.* [36] assigned known disease genes as seed nodes and computed the shortest path length between these and other nodes in the network. A node that has close proximity to multiple seed nodes receives a higher score as a candidate disease gene. The authors evaluated the method in predicting genes associated with Alzheimer's disease and showed that the genes predicted by their approach agreed with the manually curated candidates.

Neither of the above two local distance measurements captures the overall interaction network structure. As demonstrated by Kohler *et al.* [38], the closeness of two proteins cannot be fully represented by their shortest path length. Different network structures surrounding two proteins (eg. two proteins are connected by a hub, or by a protein with a low degree, or through more than one shortest path) imply different degrees of closeness between them. Global distance measurements can catch that difference by allowing equal probability of one protein to diffuse along the links of the PPI network. They tested 110 disease families containing 783 genes in prioritizing disease genes using local distance measures (direct neighborhood and shortest path length) and global similarity measures (diffusion kernel and random walk with restart). The random walk with restart method achieved an area under Receiver Operating Characteristic curve of up to 98% on simulated linkage intervals containing 100 genes, the best performance among all of the tested methods. The other global similarity-based diffusion kernel approach is also superior to the local distance measurement methods, although its performance is slightly poorer than random walk with restart.

Navlakha and Kingsford [39] compared the performance of disease gene prediction using different distance measurement methods including network neighbors, random walk with restart, propagation flow, unsupervised graph partitioning, Markov clustering and semi-supervised graph partitioning. They obtained unweighted PPI networks from HPRD and the Online Predicted Human Interaction Database (OPHID), grouped diseases from the OMIM morbid-map file based on their names and extracted loci for the associated genes from UniProt [40, 41]. They reported that random walk with restart gave the best performance in terms of precision and recall, while both random walk and propagation flow are superior to clustering and neighborhood methods. They showed that each of these methods made novel predictions that were not uncovered by another, and that only a few incorrect predictions were made using the combined methods. Hence a consensus method combining all 13 closeness measurements was proposed and selected in tandem for the ensemble of decision trees using a random forest classifier. It was demonstrated that the consensus method gave the best performance due to its ability to capture different topological properties of the PPI network.

Integration of large-scale genomic data

Many integrative approaches have been proposed for uncovering disease genes based on the assumptions that the disease genes would also share common features in gene ontology annotations, gene expression, protein sequences and domains and are likely involved in similar biological pathways and functional pathways [8]. While better prediction performance can often be achieved by integrating multiple data sources [42], the question lies in how to incorporate these heterogeneous data together for learning.

Endeavour, a prioritization algorithm through genomic data fusion, integrated more than 10 features and ranked the candidate genes based on their similarity to known disease genes for each of these features [43]. The authors first collected information for known disease genes by considering functional annotations, microarray expression, EST expression, literature, protein domains, PPIs, pathway membership, *cis*-regulatory modules, transcriptional motifs, sequence similarity and other potential data sources to be added by users. Then, candidate genes of interest were ranked based on their similarity to known disease genes in each of these features. A global ranking to prioritize candidate genes was generated by combining the ranks of individual features using order statistics. Not surprisingly, the performance based on all the data sources was shown to be much better than using partial data sources. The correct gene, in the validation of 703 disease and pathway genes, was ranked 10th among 100 candidate genes on average.

Functionally linked networks were proposed for prioritizing candidate genes by consolidating information from various data sources using a Bayesian classifier [44, 45]. Prioritizer first constructed four types of functional networks by combining different types of data sources such as gene ontology, gene expression and PPI [44]. Artificial susceptibility loci containing 50–150 genes, in steps of 50, surrounding the known disease genes were generated. The closeness in the functional network of a candidate gene in one susceptible locus to genes residing in another locus was assessed and assigned a higher score for a shorter distance. A permutation test was performed to generate *P*-values for prioritizing candidate genes. Prioritizer reached 2.8-fold enrichment compared to random selection.

While at least two susceptible loci are desired in Prioritizer, Linghu *et al.* [45] performed genome-wide prioritization by constructing an evidence-weighted functional linkage network of 21 657 genes based on 16 data sources. Pair-wise functional associations among genes in each feature were integrated into a single functional linkage network, weighted by overall functional associations, using a naïve Bayes classifier. For any given disease, scores of candidate genes for prioritization were assigned based on the sum of the weights of the network links to known disease genes. The algorithm was tested on prioritizing disease genes for 110 diseases using gene-centric and disease-centric assessments and showed outstanding performance. By testing the monogenic, polygenic and cancer disease families grouped by Kohler *et al.* [38] based on similar phenotypes, the authors observed the best performance for monogenic disease families. The fact that the performance using the integrated functional network (62% success rate) is better than using the PPI network alone (40% success rate) confirms the importance of data integration for prioritizing candidate disease genes.

Integration of phenotypic information

It has been shown that diseases with similar phenotypes often share either a common set of underlying genes or functionally related genes [46]. This observation was used to construct disease networks in which two diseases are connected, if they share at least one common gene [8]. A number of different approaches have been developed to score similarities between diseases. Rzhetsky *et al.* performed a study on 1.5 million patient records and 161 disorders using a statistical model and found that disease phenotypes form a highly connected network with strong pair-wise correlations [47]. A similar disease phenotype network was constructed by connecting diseases based on their co-occurrences in a large number of patients [48]. A couple of text mining techniques were used to map OMIM diseases to different standard vocabularies, Medical Subject Headings (MeSH) or the Unified Medical Language System (UMLS), to score pair-wise similarities among the disease records [49, 50]. Other phenotype similarity measures were also reported based on reciprocal references or the constructed human phenotype ontology [51, 52]. The scores from the disease phenotype network have been indicated to be positively correlated with several measures of gene functions [49]. A particular example where interactome and phenotype networks can reinforce each other was shown for spinocerebellar ataxia. Lim *et al.* [53] and Kahle *et al.* [54] used Medicare patient records to determine if any disease associated with proteins in the ataxia interactome also co-occurs with hereditary ataxia. One of the diseases that comorbid with ataxia was macular degeneration (MD). The ataxia interactome is significantly ($P = 7.37e\text{-}5$) enriched with proteins that interact with known MD-causing proteins, forming a MD subnetwork.

Based on the assumption that phenotypically overlapping diseases share functionally similar underlying genes, it is desirable to incorporate such phenotypic similarity profiles to candidate gene prioritization. Several studies reported that the integration of disease phenotype networks and PPI networks outperform other approaches in the prioritization task [50, 51, 55–58]. Wu *et al.* [55] used a simple linear regression method called CIPHER (Correlating protein Interaction network and PHEnotype network to pREdict disease genes) to model the correlation between phenotype similarity profile and gene closeness profile in the PPI network. The underlying assumption of the algorithm is that the phenotype similarity between two diseases can be explained by the proximity of the disease genes in the PPI network. The authors obtained the phenotype similarity data from van Driel *et al.*'s [49] text mining results and generated the network of 72 431 unique pair-wise binary interactions between 14 433 human genes by combining manually curated PPIs from HPRD, BIND, MINT and predicted PPIs from OPHID. The Pearson correlation

coefficient of the disease similarity profile for disease p and the gene closeness profile for gene g is calculated using the proposed linear regression model and recorded as a concordance score to represent the association of gene g and disease p. They showed that their predictions are reliable in prioritizing candidate genes in both linkage intervals and the entire genome, and more importantly, can potentially be applied to gene discovery for diseases without any known associated genes. Further, they demonstrated that the performance of CIPHER is comparable to that of Endeavor, an integrative approach that employed more than 10 large-scale genomic data as discussed above [43]. Interestingly, the authors showed that the direct neighbor approach for measuring the proximity of genes in the PPI network outperforms the shortest path length approach. However, as the authors addressed, the direct neighbor approach failed to assign ranks to many novel susceptibility genes in a breast cancer case study.

A similar approach was developed by Vanunu *et al.* [58] who adopted the same phenotype similarity metric computed by van Driel *et al.* [49]. They calculated the association between a query disease d and a protein p with a known disease gene for another disease d' using a logistic function dependent on the phenotype similarity between d and d'. This disease protein association was then used as prior knowledge in the constructed prioritization function, representing the relevance of protein p with disease d, to iteratively smooth itself over the network using the network propagation formula (Table 1). This algorithm, named PRINCE (PRIoritizatioN and Complex Elucidation), was demonstrated to successfully predict not only genes, but also protein complexes associated with a disease. In addition to the utilization of weighted (PRINCE) and unweighted (CIPHER) PPI networks, the major difference for PRINCE and CIPHER is that PRINCE utilized a global network propagation approach, while CIPHER only used local distance measure approaches [55]. Not surprisingly, PRINCE showed superior performance over CIPHER in prioritizing genes for 1369 diseases with a known causal gene by ~10% in ranking the real disease gene as the top-scoring one. The authors also showed that their approach outperforms the random walk with restart method [38]. Interestingly, the opposite conclusion was drawn by Navlakha and Kingsford [39] as discussed earlier. This discrepancy indicates the performance difference between random walk with restart and propagation flow might be marginal and fluctuate with different data sources and network setup.

Li and Patra constructed a heterogeneous network by integrating the PPI network and phenotype network based on disease–gene relationships in the OMIM [57]. The authors developed a new algorithm by extending the random walk with restart algorithm from only the PPI network to the entire heterogeneous network. The random walker is no longer restricted in the gene network but is also allowed to jump to the phenotype network. This Random Walk with Restart on Heterogeneous network (RWRH) algorithm prioritizes the genes and phenotypes simultaneously. In comparison with CIPHER, it showed that RWRH was superior in prioritizing disease genes under three different circumstances: known disease genes and genetic loci, known disease genes but no known genetic loci and no known disease genes or loci [55]. Further, RWRH was demonstrated to outperform random walk with restart with Area Under Curve (AUC) values of 0.96 and 0.92 respectively in prioritizing disease genes [38]. The inclusion of a phenotype network and the improved algorithm in smoothing both molecular and phenotype networks greatly enhanced the disease gene prioritization performance.

Other biological information has also been combined into the gene-phenotype heterogeneous network to aid in finding disease genes. Based on the hypothesis that disease genes and their interaction partners should have more deleterious single nucleotide polymorphisms (SNPs) than other genes, Care *et al.* [51] predicted deleterious SNPs using the random forest classifier and incorporated this information along with the PPI and phenotype networks for predicting disease genes using the same classifier. The predicted deleterious SNPs were higher in disease genes, and the inclusion of such information increased the average recall by 4% based on all PPI data and 1% based on PPI from high-throughput experiments.

Construction of disease modules

In addition to the global candidate gene prioritization algorithms, significant efforts have been made towards the discovery of disease genes for individual diseases by constructing disease modules [10]. Network components in such topological modules are believed to be functionally related and the

breakdown of one module will result in a particular disease. The information for known disease genes are collected and used to construct disease modules or subnetworks, in which members would share similar functions, expression patterns or metabolic pathways. The concept has been employed in the study of various diseases including, but not limited to, different types of cancers, type 2 diabetes, obesity, asthma, neurological diseases and so on [59–63]. This disease module approach, especially for not well-studied diseases, often requires major experimental efforts to identify interactions for constructing the module of interest.

Liu *et al.* [62] used a network-based approach and identified an insulin signaling module as well as a network of nuclear receptors that play significant roles in type 2 diabetes. Together with a subnetwork of PPIs, the authors suggested the underlying biological processes for this disorder. In a study of obesity, tissue–tissue co-expression networks between genes in the hypothalamus, liver or adipose tissue were constructed and enabled the identification of disease-specific genes [61]. The study showed that many genes included in the subnetworks were involved in obesity-related biological functions such as circadian rhythm, energy balance, stress response or immune response.

A slightly different approach was developed to prioritize disease-specific genes by constructing disease- and condition-specific subnetworks [64]. Disease-specific genes, such as differentially expressed genes identified under disease conditions, were mapped to global PPI network. The shortest path subnetwork was then built by including only the nodes in the shortest path connecting the disease-specific genes. Each node in this subnetwork was evaluated and assigned a topological score by comparing the number of shortest paths of node pairs traversing it in this subnetwork to the number of shortest paths through it in the global network. This topological scoring algorithm was verified using gene expression data from psoriasis patients and was able to identify novel targets of psoriasis.

FUTURE PERSPECTIVES

In summary, enormous progress has been made towards decoding the molecular networks and predicting novel genes associated with diseases based on these networks (Figure 2). Various distance measurements of two gene products in a network were explored and the global distance methods were demonstrated to be superior to local distance measurements. In the postgenomic era, with the deluge of large-scale genomic data, integrative approaches have also been developed to combine these types of data for prioritizing candidate genes for human diseases. However, these integrative methods tend to use overly simple network distance measurements. As shown by several groups, the inclusion of phenotype similarity networks significantly increases the performance in prioritizing disease genes. In certain cases, approaches combining phenotype networks alone even outperform those combining other multiple types of data. The utilization of phenotype similarities can potentially be used for *ab initio* predictions where no known genes are identified for certain diseases. Nevertheless, caution needs to be taken for obtaining such similarity scores between diseases based on text mining as biases and circularity can be induced [65], which will lead to an overestimation of the performance.

Since a direct comparison of different methods is often difficult due to the unavailability of some algorithms and the usage of different data sources, self-reported performance such as fold enrichment was sometimes used for comparison. One major caveat of such comparisons is that the differences in performance might be due to differences in the input data sets rather than the algorithms themselves. As shown earlier, different input interaction data sets can lead to very different performances [38]. In most prioritization methods, all known disease genes were considered equally. It might be useful to develop new algorithms to assign different weights to known disease genes in finding novel ones. Furthermore, distinct effects of node removal (complete loss of gene products) and edgetic perturbations (edge-specific interruptions) to the molecular networks should be recognized to confer different functional consequences [66]. The incorporation of such distinct perturbations should significantly improve the specificity in prioritizing candidate genes. Although not discussed in this review, other non-network-based methods for prioritizing disease genes should also be appreciated [67–69].

Although huge efforts have been made toward finding PPIs in human, we still have an incomplete map of the network due to the high complexity of networks. Common problems for predicting disease genes based on networks are the existence of noise (false positives) in curated PPI databases and gene

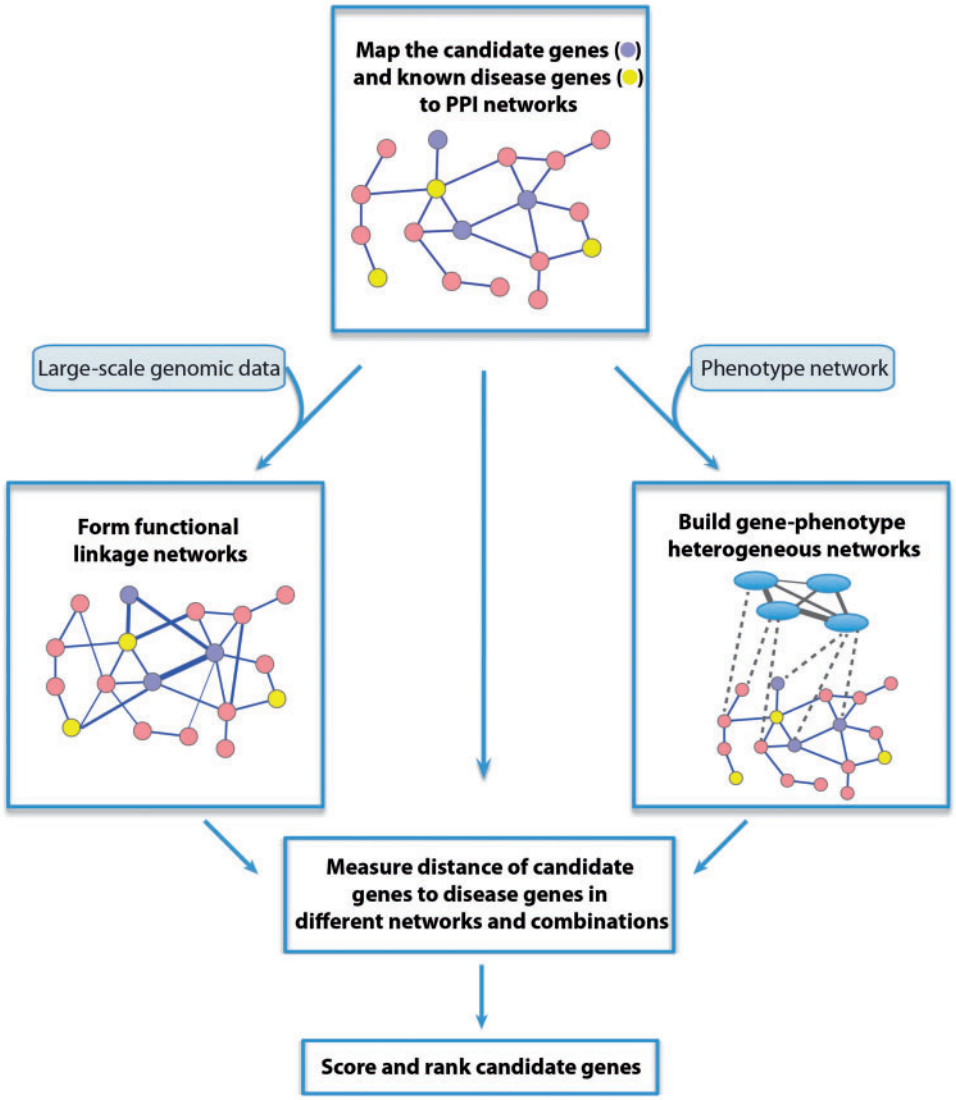

Figure 2: Prioritizing schemes for finding disease-associated genes. Candidate genes (within linkage intervals or genome wide) and known disease genes are mapped to interactome networks. The distance between candidate genes and known disease genes in interactome networks, functional networks or gene-phenotype networks are measured using different methods to score and rank candidate genes.

expression profiling experiments in addition to the bias towards well-studied disease genes [23, 70]. High quality molecular networks are desired to increase the prediction power and are realizable with advances in high-throughput methods [26]. While efforts have been made mostly on human molecular networks, it is worth noticing that an increasing number of protein interaction networks are under construction for microbial pathogens [71–76]. Combining viral protein networks and human protein networks, so called 'virhostome', might unravel key mechanisms of pathogen infection since virus–host interactions are mostly physical interactions [11].

To conclude, the integration of steadily growing cellular interactomes including PPI networks, regulatory networks, metabolic networks and virus–host networks are crucial for understanding the mechanisms of human diseases and predicting novel candidate genes associated with diseases.

Key Points

- Human diseases are the consequences of disruption in molecular networks.
- Genes associated with the same or similar diseases tend to reside in the same neighborhood of molecular networks.

- Network-based computational approaches have been developed to find novel disease genes and prioritize candidate genes.
- Global distance measurements between candidate genes and known disease genes in networks outperform local distance measurement approaches in prioritizing candidate genes.
- The integration of large-scale genomic data or phenotypic information with networks greatly increases the prediction performance.

FUNDING

This work is supported by start-up funds to H.Y. from Cornell University.

Acknowledgements

We thank Siu Sylvia Lee and Elliot John Kahen for their valuable comments on the manuscript.

References

1. Kann MG. Advances in translational bioinformatics: computational approaches for the hunting of disease genes. *Brief Bioinform* 2010;**11**:96–110.
2. Oldenburg RA, Meijers-Heijboer H, Cornelisse CJ, *et al*. Genetic susceptibility for breast cancer: how many more genes to be found? *Crit Rev Oncol Hematol* 2007;**63**:125–49.
3. Frayling TM. Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat Rev Genet* 2007;**8**:657–62.
4. Zhu M, Zhao S. Candidate gene identification approach: progress and challenges. *Int J Biol Sci* 2007;**3**:420–7.
5. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. *Science* 2008;**322**:881–8.
6. Vidal M. A unifying view of 21st century systems biology. *FEBS Lett* 2009;**583**:3891–4.
7. Oti M, Brunner HG. The modular nature of genetic diseases. *Clin Genet* 2007;**71**:1–11.
8. Goh KI, Cusick ME, Valle D, *et al*. The human disease network. *Proc Natl Acad Sci USA* 2007;**104**:8685–90.
9. Feldman I, Rzhetsky A, Vitkup D. Network properties of genes harboring inherited disease mutations. *Proc Natl Acad Sci USA* 2008;**105**:4323–8.
10. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011;**12**:56–68.
11. Vidal M, Cusick ME, Barabasi AL. Interactome networks and human disease. *Cell* 2011;**144**:986–98.
12. Seebacher J, Gavin AC. SnapShot: Protein-protein interaction networks. *Cell* 2011;**144**:1000.
13. Rual JF, Venkatesan K, Hao T, *et al*. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005;**437**:1173–8.
14. Stelzl U, Worm U, Lalowski M, *et al*. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 2005;**122**:957–68.
15. Yu H, Tardivo L, Tam S, *et al*. Next-generation sequencing to generate interactome datasets. *Nat Methods* 2011;**8**:478–80.
16. Charbonnier S, Gallego O, Gavin AC. The social network of a cell: recent advances in interactome mapping. *Biotechnol Annu Rev* 2008;**14**:1–28.
17. Ewing RM, Chu P, Elisma F, *et al*. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol* 2007;**3**:89.
18. Peri S, Navarro JD, Amanchy R, *et al*. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* 2003;**13**:2363–71.
19. Chatr-aryamontri A, Ceol A, Palazzi LM, *et al*. MINT: the Molecular INTeraction database. *Nucleic Acids Res* 2007;**35**:D572–4.
20. Stark C, Breitkreutz BJ, Reguly T, *et al*. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006;**34**:D535–9.
21. Bader GD, Donaldson I, Wolting C, *et al*. BIND–The Biomolecular Interaction Network Database. *Nucleic Acids Res* 2001;**29**:242–5.
22. Kerrien S, Alam-Faruque Y, Aranda B, *et al*. IntAct–open source resource for molecular interaction data. *Nucleic Acids Res* 2007;**35**:D561–5.
23. Cusick ME, Yu H, Smolyar A, *et al*. Literature-curated protein interaction datasets. *Nat Methods* 2009;**6**:39–46.
24. Jeong H, Mason SP, Barabasi AL, *et al*. Lethality and centrality in protein networks. *Nature* 2001;**411**:41–2.
25. Ivanic J, Yu X, Wallqvist A, *et al*. Influence of protein abundance on high-throughput protein-protein interaction detection. *PLoS One* 2009;**4**:e5815.
26. Yu H, Braun P, Yildirim MA, *et al*. High-quality binary protein interaction map of the yeast interactome network. *Science* 2008;**322**:104–10.
27. Yu H, Kim PM, Sprecher E, *et al*. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* 2007;**3**:e59.
28. Yu H, Zhu X, Greenbaum D, *et al*. TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res* 2004;**32**:328–37.
29. Wachi S, Yoneda K, Wu R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* 2005;**21**:4205–8.
30. Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. *Bioinformatics* 2006;**22**:2291–7.
31. Oliver S. Guilt-by-association goes global. *Nature* 2000;**403**:601–3.
32. Altshuler D, Daly M, Kruglyak L. Guilt by association. *Nat Genet* 2000;**26**:135–7.
33. Wang PI, Marcotte EM. It's the machine that matters: Predicting gene function and phenotype from protein networks. *J Proteomics* 2010;**73**:2277–89.
34. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. *Mol Syst Biol* 2007;**3**:88.
35. Oti M, Snel B, Huynen MA, *et al*. Predicting disease genes using protein-protein interactions. *J Med Genet* 2006;**43**:691–8.
36. Krauthammer M, Kaufmann CA, Gilliam TC, *et al*. Molecular triangulation: bridging linkage and molecular-

network information for identifying candidate genes in Alzheimer's disease. *Proc Natl Acad Sci USA* 2004;**101**: 15148–53.

37. Radivojac P, Peng K, Clark WT, *et al*. An integrated approach to inferring gene-disease associations in humans. *Proteins* 2008;**72**:1030–7.
38. Kohler S, Bauer S, Horn D, *et al*. Walking the interactome for prioritization of candidate disease genes. *AmJ Hum Genet* 2008;**82**:949–58.
39. Navlakha S, Kingsford C. The power of protein interaction networks for associating genes with diseases. *Bioinformatics* 2010;**26**:1057–63.
40. Brown KR, Jurisica I. Online predicted human interaction database. *Bioinformatics* 2005;**21**:2076–82.
41. Jain E, Bairoch A, Duvaud S, *et al*. Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* 2009;**10**:136.
42. Jansen R, Yu H, Greenbaum D, *et al*. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 2003;**302**:449–53.
43. Aerts S, Lambrechts D, Maity S, *et al*. Gene prioritization through genomic data fusion. *Nat Biotechnol* 2006;**24**: 537–44.
44. Franke L, van Bakel H, Fokkens L, *et al*. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* 2006;**78**:1011–25.
45. Linghu B, Snitkin ES, Hu Z, *et al*. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biol* 2009;**10**:R91.
46. Brunner HG, van Driel MA. From syndrome families to functional genomics. *Nat Rev Genet* 2004;**5**:545–51.
47. Rzhetsky A, Wajngurt D, Park N, *et al*. Probing genetic overlap among complex human phenotypes. *Proc Natl Acad Sci USA* 2007;**104**:11694–9.
48. Hidalgo CA, Blumm N, Barabasi AL, *et al*. A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol* 2009;**5**:e1000353.
49. van Driel MA, Bruggeman J, Vriend G, *et al*. A text-mining analysis of the human phenome. *Eur J Hum Genet* 2006;**14**: 535–42.
50. Lage K, Karlberg EO, Storling ZM, *et al*. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 2007;**25**:309–16.
51. Care MA, Bradford JR, Needham CJ, *et al*. Combining the interactome and deleterious SNP predictions to improve disease gene identification. *Hum Mutat* 2009;**30**: 485–92.
52. Robinson PN, Kohler S, Bauer S, *et al*. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *AmJ Hum Genet* 2008;**83**:610–5.
53. Lim J, Hao T, Shaw C, *et al*. A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* 2006;**125**:801–14.
54. Kahle JJ, Gulbahce N, Shaw CA, *et al*. Comparison of an expanded ataxia interactome with patient medical records reveals a relationship between macular degeneration and ataxia. *Hum Mol Genet* 2011;**20**:510–27.
55. Wu X, Jiang R, Zhang MQ, *et al*. Network-based global inference of human disease genes. *Mol Syst Biol* 2008;**4**:189.
56. Wu X, Liu Q, Jiang R. Align human interactome with phenome to identify causative genes and networks underlying disease families. *Bioinformatics* 2009;**25**:98–104.
57. Li Y, Patra JC. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics* 2010;**26**:1219–24.
58. Vanunu O, Magger O, Ruppin E, *et al*. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol* 2010;**6**:e1000641.
59. Taylor IW, Linding R, Warde-Farley D, *et al*. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol* 2009;**27**:199–204.
60. Chen Y, Zhu J, Lum PY, *et al*. Variations in DNA elucidate molecular networks that cause disease. *Nature* 2008;**452**: 429–35.
61. Dobrin R, Zhu J, Molony C, *et al*. Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome Biol* 2009;**10**:R55.
62. Liu M, Liberzon A, Kong SW, *et al*. Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genet* 2007;**3**:e96.
63. Hwang S, Son SW, Kim SC, *et al*. A protein interaction network associated with asthma. *J Theor Biol* 2008;**252**: 722–31.
64. Dezso Z, Nikolsky Y, Nikolskaya T, *et al*. Identifying disease-specific genes based on their topological significance in protein networks. *BMC Syst Biol* 2009;**3**:36.
65. Wang J, Zhou X, Zhu J, *et al*. Bias of phenotype similarity scores between diseases. *International Conference on Bioinformatics and Biomedical Engineering*, Chengdu, China. IEEE 2010.
66. Zhong Q, Simonis N, Li QR, *et al*. Edgetic perturbation models of human inherited disorders. *Mol Syst Biol* 2009;**5**: 321.
67. Adie EA, Adams RR, Evans KL, *et al*. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics* 2005;**6**:55.
68. Adie EA, Adams RR, Evans KL, *et al*. SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics* 2006;**22**:773–4.
69. Gaulton KJ, Mohlke KL, Vision TJ. A computational system to select candidate genes for complex human traits. *Bioinformatics* 2007;**23**:1132–40.
70. Yu H, Nguyen K, Royce T, *et al*. Positional artifacts in microarrays: experimental verification and construction of COP, an automated detection tool. *Nucleic Acids Res* 2007; **35**:e8.
71. Calderwood MA, Venkatesan K, Xing L, *et al*. Epstein-Barr virus and virus human protein interaction maps. *Proc Natl Acad Sci USA* 2007;**104**:7606–11.
72. de Chassey B, Navratil V, Tafforeau L, *et al*. Hepatitis C virus infection protein network. *Mol Syst Biol* 2008;**4**:230.
73. Uetz P, Dong YA, Zeretzke C, *et al*. Herpes viral protein networks and their interaction with the human proteome. *Science* 2006;**311**:239–42.
74. Shapira SD, Gat-Viks I, Shum BO, *et al*. A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell* 2009;**139**:1255–67.
75. Mendez-Rios J, Uetz P. Global approaches to study protein-protein interactions among viruses and hosts. *Future Microbiol* 2010;**5**:289–301.

76. Jager S, Gulbahce N, Cimermancic P, *et al*. Purification and characterization of HIV-human protein complexes. *Methods* 2011;**53**:13–9.

77. Karni S, Soreq H, Sharan R. A network-based method for predicting disease-causing genes. *J Comput Biol* 2009;**16**: 181–9.

78. Salwinski L, Miller CS, Smith AJ, *et al*. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* 2004; **32**:D449–51.

79. von Mering C, Jensen LJ, Kuhn M, *et al*. STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* 2007;**35**:D358–62.

80. Mewes HW, Dietmann S, Frishman D, *et al*. MIPS: analysis and annotation of genome information in 2007. *Nucleic Acids Res* 2008;**36**:D196–201.

81. Kanehisa M, Goto S, Hattori M, *et al*. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006;**34**:D354–7.

82. Joshi-Tope G, Gillespie M, Vastrik I, *et al*. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 2005;**33**:D428–32.