BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 22 2014, pages 3223–3231
doi:10.1093/bioinformatics/btu482

Systems biology

Advance Access publication July 30, 2014

Network-based analysis of genotype–phenotype correlations between different inheritance modes

Dapeng Hao[1,*], Chuanxing Li[1,2], Shaojun Zhang[1], Jianping Lu[1], Yongshuai Jiang[1], Shiyuan Wang[1] and Meng Zhou[1,*]

[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, P.R. China and [2]Institute for Systems Biology, Seattle 98109, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Recent studies on human disease have revealed that aberrant interaction between proteins probably underlies a substantial number of human genetic diseases. This suggests a need to investigate disease inheritance mode using interaction, and based on which to refresh our conceptual understanding of a series of properties regarding inheritance mode of human disease.

Results: We observed a strong correlation between the number of protein interactions and the likelihood of a gene causing any dominant diseases or multiple dominant diseases, whereas no correlation was observed between protein interaction and the likelihood of a gene causing recessive diseases. We found that dominant diseases are more likely to be associated with disruption of important interactions. These suggest inheritance mode should be understood using protein interaction. We therefore reviewed the previous studies and refined an interaction model of inheritance mode, and then confirmed that this model is largely reasonable using new evidences. With these findings, we found that the inheritance mode of human genetic diseases can be predicted using protein interaction. By integrating the systems biology perspectives with the classical disease genetics paradigm, our study provides some new insights into genotype–phenotype correlations.

Contact: haodapeng@ems.hrbmu.edu.cn or biofomeng@hotmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 21, 2013; revised on June 22, 2014; accepted on July 8, 2014

1 INTRODUCTION

Decades of research into human genetic diseases, revolutionized by the application of innovative technologies including next-generation sequencing and genome-wide association study, has led to the accumulation of an impressive amount of disease–gene associations (Broeckel and Schork, 2004; Handel *et al.*, 2013; Ku *et al.*, 2011). On the other hand, the rapid development of proteomic technologies provides an extensive protein interaction map that improves our understanding of the complex genotype-to-phenotype relationships among human diseases and their associated genes (Rual *et al.*, 2005; Stelzl *et al.*, 2005). This provides an opportunity to reevaluate the traditional models about human genetic diseases. An important molecular mechanism that underlies genotype-to-phenotype relationship is the inheritance modes, including either dominant mode that a single mutated allele of relevant gene is enough to affect or recessive mode that both mutated alleles are necessary. Although the rapid accumulation of disease–gene associations and protein interaction data promise to improve our understanding of the disease inheritance mode and how it interrelates to the human diseases at a systems level, the incorporation of inheritance mode has been rarely seen in recent large-scale studies.

Protein network-based study has been proven to be a valuable strategy for understanding the molecular mechanism underlying human diseases (Barabasi *et al.*, 2011; Goh *et al.*, 2007). Recently, the network-perturbation model has been proposed as an alternative molecular mechanism of human diseases (Furlong, 2013; Muers, 2010; Walhout, 2009), arguing that the inheritance mode should also be recognized at the network level. In this model, failures in the connectivity of protein interaction network (PIN) that alter the network topology underlie human diseases. It was observed initially in *Caenorhabditis elegans* that mutations of gene RbAp48 can cause the loss of a subset of interactions while leaving the gene partially functional (Walhout *et al.*, 2000). Subsequently, a number of studies have provided strong evidence that mutations causing specific loss of interaction, not necessarily complete loss of gene product, frequently occur in diseases (Schuster-Bockler and Bateman, 2008). Many disorders, such as Charcot Marie Tooth disease, Creutzfeldt Jacob disease and Alzheimer syndrome, were found to be protein interaction-related diseases (Chiti and Dobson, 2006; Ross *et al.*, 2005; Shy *et al.*, 2004). Researchers also suggested that aberrant protein interactions involving gene Htt may contribute to the pathology of Huntington (Cattaneo *et al.*, 2001; Giorgini and Muchowski, 2005; Zhang *et al.*, 2003). A recent study demonstrated that in-frame Mendelian disease mutations are highly enriched on protein interaction interfaces, suggesting that protein interaction plays a role in the pathogenesis of a substantial number of genetic diseases (Wang *et al.*, 2012).

So far, the classical model of genotype-to-phenotype relationship is gene-centric, assuming that a mutation causes complete loss of a gene product (Muers, 2010). However, the network-perturbation model suggests that, in principle, the understanding of Mendelian inheritance should take into account of the interaction between two genes. To our knowledge, two recent studies analyzed the connection between inheritance modes of human genetic diseases and protein interactions (Schuster-Bockler and

*To whom correspondence should be addressed.

Bateman, 2008; Zhong *et al.*, 2009). Schuster-Bockler and Bateman collected 119 known protein interaction-altering mutations from scientific literatures and observed a significant enrichment of autosomal dominant (AD) mutations versus autosomal recessive (AR) mutations. Zhong *et al.* found that AD disorders are more frequently associated with in-frame mutations than AR disorders, a mutation type that frequently leads to defective protein interaction rather than complete gene loss. Given the previous findings about the correlation between interaction and inheritance mode, it is highly probable that protein network properties can differentiate between different inheritance modes. Here, by taking advantage of comprehensive protein interaction data and a large number of Mendelian diseases with known inheritance mode, we conducted a systematic study to show the correlation between protein network and inheritance modes, and refined a likely interaction-based model to explain the correlation.

2 MATERIALS AND METHODS

2.1 Disease mutations, disease genes and inheritance modes

We obtained inheritance modes, disorder–gene associations and disease mutations from the Online Mendelian Inheritance in Man (OMIM) database (Hamosh *et al.*, 2005) and M2SG database (Ji *et al.*, 2013). A schematic illustration for the data extraction and integration from OMIM was provided in Supplementary Figure S1. First, the inheritance mode of diseases was extracted from OMIM for 4409 disease entries. Diseases with unclear inheritance and multiple inheritance modes were not included. Then, the inheritance mode of mutations or disease genes was determined by the inheritance mode of associated diseases: 1234 AD genes were defined if single mutated allele of the gene is enough to affect disease phenotypes, and 1162 AR genes were defined if both mutated alleles of the gene are necessary to cause any disease phenotypes. According to this definition, 181 AD genes also contain mutations causing AR disease. AD interactions and AR interactions in our manuscript are referred to as interactions between AD genes and AR genes, respectively.

All mutations and affected phenotypes were extracted independently from OMIM (Supplementary Table S1). Only mutations mapped into the diseases in diseases were used. Inheritance mode of OMIM mutations and M2SG mutations was determined by manually comparing their disease names with OMIM disease entries. A mutation pair is the combination of two disease mutations, and all possible combinations of mutations were considered when analyzing the mutation pairs. Disease mutations were divided into three classes: in-frame mutations (including missense mutations and small in-frame indels), out-frame mutations (including nonsense mutations and frameshift indels) and other mutations (i.e. fusions).

Disease–gene associations were provided with inheritance in Supplementary Table S2. To integrate different types of disease information, 12 465 disease names were manually merged into 2463 diseases based on their given names (Supplementary Table S3). Pleiotropic gene was defined if the gene was associated with multiple diseases based on our classification.

2.2 Protein interaction network

The human protein binary interactions were compiled from following resources on June 2012: MINT (Licata *et al.*, 2012); BioGRID (Chatr-Aryamontri *et al.*, 2013); IntAct (Kerrien *et al.*, 2012); DIP (Salwinski *et al.*, 2004); BIND (Isserlin *et al.*, 2011); HPRD (Keshava Prasad *et al.*, 2009); iRefWeb 4.1 to integrate the interactions from innatedb, matrixdb and mppi (Turner *et al.*, 2010); four datasets from public publications (Rual *et al.*, 2005; Stelzl *et al.*, 2005; Venkatesan *et al.*, 2009; Yu *et al.*, 2011). Experimental methods used to generate the interaction were checked to filter out the methods only detecting protein complexes. Predicted interactions were not considered. Finally, we constructed a reliable PIN comprising 77 192 interactions between 12 869 human genes.

A total of 3347protein complexes were compiled from Corum database release of February 2012 (Ruepp *et al.*, 2010) and HPRD release 9 (Keshava Prasad *et al.*, 2009).

Hub genes and bottleneck genes were defined by top 20% genes sorted by node degree and betweenness, respectively.

2.3 Structurally resolved protein interaction interfaces

The interfaces of protein interactions were structurally resolved by a homology modeling approach, as described in the work of Wang *et al.* (2012). Pfam (version 26.0), 3did (version 'Apr_3_2011') and iPfam data (September 3, 2012) were used to identify the interfaces of two interacting proteins (Kelley and Ideker, 2005; Yu *et al.*, 2007). We are aware that a rigorous analysis of the interaction interface in interacting proteins should be done based on a full atom-level description of the complete structure of protein complex. In the absence of this information and knowing it is not a general rule, the interacting domains on the corresponding interacting proteins were considered to be the interaction interface of the interaction. In total, 6823 interacting domain pairs of 4438 individual domains were used. Protein domain information was taken from UniProt and Pfam. To map disease mutations into protein domains, we only considered mutations for which the codon information was available, and then identified protein domains using Pfam and UniProt from the codon information, as described in the work of Park *et al.* (2009).

2.4 Detecting network modules

A probabilistic modeling algorithm was used to capture the modules of PIN (Kelley and Ideker, 2005). The algorithm uses a log-odds score to assess the likelihood that a set of nodes is more densely interconnected than expected by chance:

$$OS = \log \frac{\prod_{a,b \in M} \beta I_E(a,b) + (1-\beta)(1-I_E(a,b))}{\prod_{a,b \in M} r_{a,b} I_E(a,b) + (1-r_{a,b})(1-I_E(a,b))}$$

Where M is a putative module of proteins and E is a set of interactions inside the module. $I_E(a,b) = 1$ if there is an interaction between members a and b and otherwise $I_E(a,b) = 0$. For a densely interconnected module M, members are expected to interact with each other with a high probability β, which is higher than the random probability $r_{a,b}$.β was set to 0.8 according to empirical study. $r_{a,b}$ was determined by estimating the probability of observing the same interaction in random networks. Significant modules were searched in PIN by a greedy search procedure, and were estimated by performing 100 random trials in random networks generated by edge-swapping method.

2.5 Network-based classification

To test the performance of large-scale properties in determining the inheritance modes of disease genes, we used three direct topological properties as classification features, including node degree, betweenness and the fraction of interactors with particular inheritance modes. To measure the diagnostic value of these features, we designed a RBF kernel-based support vector classifier (SVM) to predict the inheritance modes, and then used a 5-fold cross-validation strategy to evaluate the performance of SVM classifier. The outcome of prediction was then assessed by receiver operating characteristic curve analysis. A benchmark dataset was selected with two following criteria: (i) both genes in our PIN are AD or AR disease genes; (ii) the associated diseases have at least two disease genes.

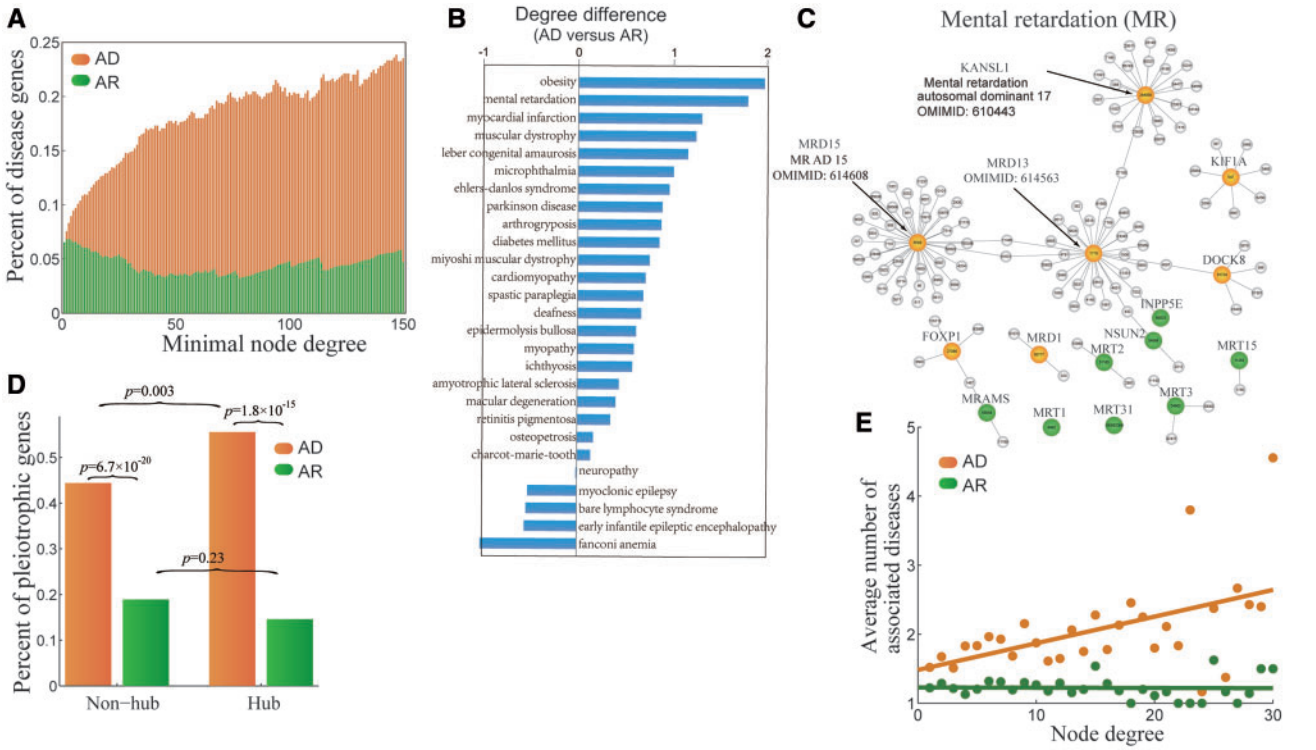

Fig. 1. Analysis of network connectivity, gene pleiotropy and their association with inheritance modes. (**A**) The probability of genes associated with AD/AR diseases as a function of network connectivity. (**B**) Network connectivity difference between dominant and recessive diseases of the same disease class. The average degree of the dominant disease-associated genes minus the average degree of recessive disease-associated genes was normalized by the average degree of the associated disease and was plotted for each of the diseases, as indicated. Note that to get robust result only disease classes with at least three associated genes for each inheritance mode are analyzed. (**C**) Protein interactions of disease genes associated with mental retardation, where genes are labeled by Entrez ID. Disease genes of unknown inheritance mode are not included. (**D**) The fraction of pleiotropic genes of hub genes and non-hub genes with respect to inheritance modes (χ^2 test). (**E**) Average number of associated diseases is plotted as a function of gene connectivity in PIN, measured for AD and AR genes, respectively

2.6 Tissue homogeneity

We used global gene expression data across 79 different human tissue types from the GNF Atlas project (Su *et al.*, 2004). The tissue-specificity measure (SPM) of a gene x in a tissue i (x_i, $i \in [1 \sim 79]$) was detected by solving a linear algebra problem of scalar projection (Xiao *et al.*, 2010). SPM ranges from 0 to 1. A value (x_i) close to 1 indicates that the gene x is mainly expressed in tissue i. Then, the tissue homogeneity between genes was estimated by the Pearson correlation of SPM values across 79 tissues. If the expression of the two genes was highly selective in several similar tissues, the Pearson correlation coefficient would be close to 1, or otherwise would be close to 0.

3 RESULTS

3.1 PIN and its association with AD diseases and AR diseases

We first calculated the likelihood of genes causing any AD disease as a function of increasing node degree in PIN. This likelihood showed a dramatic increase for genes with a larger number of protein interactions, whereas no significant correlation was found for AR disease genes (Fig. 1A). We then asked whether the correlation between AD disease and network connectivity was attributed to the fact that many AD diseases are associated with genes encoding structural proteins and transcription factors, which tend to have many protein interactions. We ruled out this possibility by finding out that the trend holds even after controlling for structural proteins or transcription factors (Supplementary Fig. S2).

As a result, AD disease genes have significantly higher node degree than that of non-disease genes (Mann–Whitney U-test, $P = 10^{-54}$). However, no significant difference was found for the node degree between AR disease genes and non-disease genes (Mann–Whitney U-test, $P = 0.86$), indicating that AR disease genes have no tendency to encode hub genes or non-hub genes. We asked whether the correlation was attributed to the bias that AD genes are often well-known disease genes and thus overrepresented in PIN dataset. We thus compared the correlation between protein interaction and disease genes that are associated with the dominant inheritance and recessive inheritance of the same disease class. We found that, even for the same disease classification, the disease genes associated with a dominant disease entry have significantly higher node degree than that of disease genes associated with a recessive disease entry (sign test, $P < 0.001$, Fig. 1B). For example, in mental retardation and Parkinson disease (Fig. 1C and Supplementary Fig. S3), all the top highly connected disease genes are associated with

a dominant disease entry, whereas the poorly connected disease genes are generally associated with a recessive disease entry. These results may help to address the recent debates about whether disease genes tend to encode hubs in PINs (Furlong, 2013).

3.2 Gene pleiotropy and its association with AD diseases and AR diseases

It has been suggested that PIN can serve as a framework to understand the pleiotropy of human disease genes (Tyler *et al.*, 2009). We investigated this by separating disease genes of different inheritance modes. We found that, whether for hub genes or non-hub genes, the proportion of AD pleiotropic disease genes was significantly higher than that of AR pleiotropic disease genes (Fig. 1D). We further observed that not only the proportion of pleiotropic genes, but also the number of disorders that a pleiotropic AD disease gene associated with is significantly higher than that of a pleiotropic AR disease gene associated with (Mann–Whitney *U*-test, $P = 1.4 \times 10^{-8}$). Some hub AD genes such as COL2A1 and FGFR2 are associated with more than a dozen of entirely different AD disorders. This result suggests that AD genes are more likely to cause different diseases than AR genes, probably because AD genes carry a larger number of interactions.

More importantly, we found that the proportion of pleiotropic genes in hub AD disease genes was significantly higher than the proportion in non-hub AD disease genes; however, we did not observe significant difference for AR disease genes (Fig. 1D). Moreover, a significant correlation was found between the average number of associated AD diseases and the node degree of the AD disease genes (Pearson correlation $r = 0.5$, $P < 0.01$, Fig. 1E), whereas no significant correlation was found between the average number of associated AR diseases and the node degree of the AR disease genes. These results suggest that protein interaction can be used as a molecular basis of pleiotropy for AD disease genes, while it may not be appropriate for explaining the pleiotropy for AR disease genes.

The pleiotropy analysis was based on our disease classification. To avoid the potential bias of our disease classification, we also tested these results using OMIM disease entry and an independent disease classification system based on HGMD mutations, respectively, both of which show the same significant results (Supplementary Fig. S4).

3.3 AD Diseases are associated with the loss of important interactions

Previous studies on human disease highlight the importance of bottleneck genes that mediate the communication between functional modules (Margadant and Sonnenberg, 2010; Shao *et al.*, 2012; Taylor *et al.*, 2009; Xu *et al.*, 2011), for which the node degree might be small but with important protein interactions whose loss disrupts the cross talk between modules. We therefore investigated the association with bottleneck genes for different disease inheritance modes. We defined bottleneck genes using a topological measure known as 'betweenness' (Taylor *et al.*, 2009). In the biological context, betweenness measures the ways that signals pass though the network and is previously used to define the bottleneck genes (Yu *et al.*, 2007). In Figure 2A, we show that bottleneck genes are more likely to be associated with AD diseases, as opposed to AR diseases. The correlation is significant even after controlling for node degree (analysis of covariance test, $P = 0.037$). We found that non-hub bottleneck genes are more likely to be associated with AD diseases than hub non-bottleneck genes, even though their node degrees are dramatically lower than that of hub non-bottleneck genes (Supplementary Fig. S5), suggesting that betweenness centrality is another topological measure indicating the likelihood of a gene being associated with AD diseases. To further investigate this, we studied a comprehensive list of human protein complexes, a typical representative of modules that extracted from widely used databases. We divided members of protein complexes into bottleneck genes and non-bottleneck genes, and found that bottleneck genes are significantly more likely to be AD disease genes than non-bottleneck genes in protein complexes (Fig. 2B). However, there is no significant difference for AR disease genes. We also used a probabilistic modeling algorithm to capture the network modules (see Section 2), which revealed 432 significant modules. We examined the overlap between the 432 significant modules and found a significant enrichment of AD disease genes in the overlap of adjacent modules, as compared with the probability of finding AD disease genes in other positions (Fisher exact test, $P < 0.01$). Figure 2C shows an example of two overlapping modules, where two AD disease genes are located at the interface. The results suggest that AD diseases are more likely to associate with the disruption of important interactions. As an example, the AD disease infantile Haemangioma is associated with mutations on gene *FLT4*, a small node degree gene that mediates between several important genes (Fig. 2D). Specifically, *FLT4* contains Pkinase Tyr domain that is responsible for the connection between hub genes GRB2, KDR and SHC1, and I-set domain that is responsible for the interaction with a non-hub gene *VEGFC*. Interestingly, in-frame disease mutations are significantly enriched on Pkinase_Tyr domain, suggesting that loss of important interactions of this gene might be associated with disease.

3.4 Understanding inheritance modes using protein interaction

To explain the results, we tentatively proposed the model of AD interaction and AR interaction as in Figure 3A, based on an assumption that most disease mutations are loss-of-function, as did in previous study (Zhong *et al.*, 2009). This model refines the conclusion of previous studies (Schuster-Bockler and Bateman, 2008; Zhong *et al.*, 2009) and helps the user to illustrate that AD genes are more likely to be associated with interaction-altering than AR genes. For simplicity, we ideally modeled AD interaction when insufficiency of the interaction affects disease phenotypes, and modeled AR interaction when complete loss of this interaction affects disease phenotypes. Apparently, according to the classical gene-centric model of inheritance mode, a disease gene associated with an AD interaction is itself dominant. In the case of AD interaction, mutations on single copy of either interacting genes that render the interaction defective are able to affect the phenotype. Conversely, in the case of AR interaction, to cause the complete loss of the interaction, mutations on both copies of the same interacting gene are necessary. Moreover,

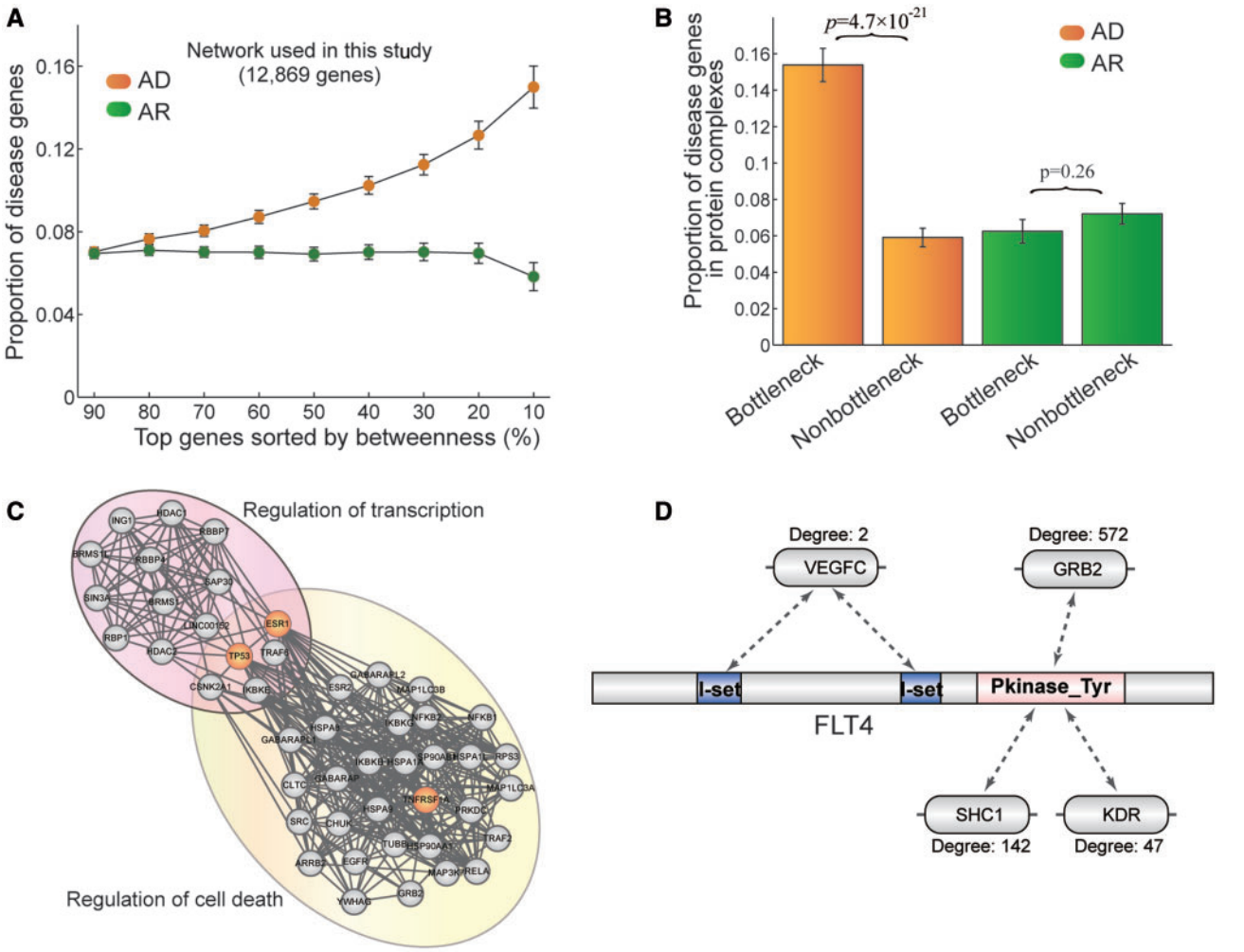

Fig. 2. Network modularity and disease inheritance modes. (**A**) Proportion of AD genes and AR genes of bottleneck genes. (**B**) Proportion of AD disease genes and AR disease genes measured for bottleneck genes and non-bottleneck genes of protein complexes. (**C**) An example of AD genes located on the overlap of two adjacent functional modules. (**D**) Illustration of FLT4 and its interactions

those mutations should cause the complete gene loss, or else mutations on both alleles should be on the same interaction interface, which is less likely to happen. Therefore, given that protein interaction underlies a substantial number of human genetic diseases, one can expect that AR diseases are more frequently associated with the complete gene loss than AD diseases. This model is consistent with previous finding that interaction-altering mutations are enriched of dominant mutations and the finding that AD disorders are more frequently associated with in-frame mutations than AR disorders (Schuster-Bockler and Bateman, 2008; Zhong *et al.*, 2009).

We did not have an appropriate dataset of mutations causing loss of interaction to confirm the model with more convincing evidences. However, we studied genes and mutations that are more likely to be associated with interaction-altering or gene loss, and analyzed their predisposition to cause AD diseases and AR diseases. Using OMIM data, we found that AR disease genes were more likely to have out-frame mutations than AD disease genes (600 versus 373 genes, χ^2 test, $P = 2 \times 10^{-10}$), strengthening the idea that AR diseases are more frequently associated with complete gene loss than AD disease genes. We also plotted the ratio of the number of AR/the number of AD genes as function of the minimal fraction of out-frame/in-frame mutations (Fig. 3B), and found that the number of AR genes was >4-fold higher than the number of AD genes if all the mutations on the disease genes are out-frame. As an example, we collected 175 AR mutations and 103 AD mutations associated with disease 'DEAFNESS', and found that AR mutations are more likely to be deleterious than AD mutations (21.1 versus 4.9% out-frame mutations, $P = 2 \times 10^{-4}$, Fisher exact test).

We then ask whether AD disease genes are more likely to contain protein domains involving in protein–protein binding than AR disease genes. For this purpose, we analyzed the promiscuous domains that present in various combinations in multi-domain proteins and participate in many different kinds of protein interactions (Basu *et al.*, 2008). We found a significantly higher frequency of finding a promiscuous domain in AD disease genes than in AR disease genes (Sign test, $P < 0.01$). For

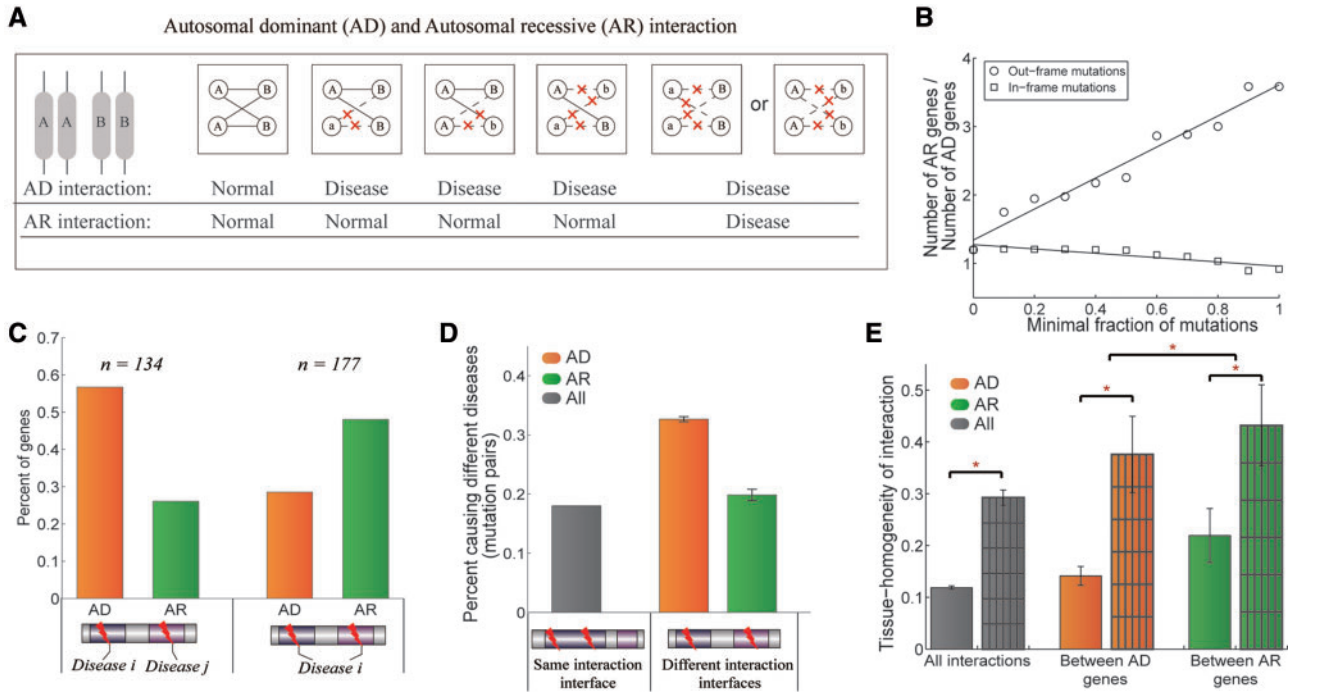

Fig. 3. Interaction model of inheritance. (**A**) The AD interaction and AR interaction. Dominant interaction is ideally defined if the insufficiency of the interaction has phenotype effect, and recessive interaction is defined if the complete loss of the interaction has phenotype effect. The X mark means a mutation on the interaction interface causing the loss of corresponding interaction. The model suggests AD interactions are easier to happen than AR interaction. (**B**) The ratio of the number of AR genes versus the number of AD genes is plotted as minimal fraction of out-frame/in-frame mutations on a gene. (C) Fraction of AD disease genes and AR disease genes with respect to the mutations on different interaction interfaces causing different diseases and the same disease. (**D**) Percentage of mutation pairs of the same protein causing different diseases. Errors are calculated by bootstrapping based on resampling. (**E**) Tissue homogeneity of protein interactions. Bars without (with) gridlines represent both genes of the interaction causing different diseases (same disease). In cases of interactions between AD/AR genes, only interactions with mutations on corresponding interaction interfaces of interacting proteins were considered. Asterisk represents statistical significance (Mann–Whitney *U*-test, $P<0.05$)

example, two promiscuous domains involving in protein–protein binding, PHD and PDZ, were found predominantly in AD disease genes. We also analyzed a list of genes with in-frame mutations significantly enriched on protein interaction interfaces (Wang *et al.*, 2012), and found that AD genes were significantly more enriched than AR genes in the list (Fisher exact test, $P = 7 \times 10^{-5}$). Then, we mapped disease mutations into the interacting interfaces of the associated genes following the previous studies (Das *et al.*, 2014; Wang *et al.*, 2012). We found in total 134 genes having mutations on different interaction interfaces causing different diseases, suggesting that different diseases are probably associated with different interactions (Fig. 3C). Among these genes, the proportion of AD disease genes was~2-fold higher than that of AR disease genes (Fisher exact test, $P = 4 \times 10^{-4}$). As a comparison, we analyzed 177 genes having mutations on different interfaces causing the same disease. Conversely, among the 177 genes, the proportion of AD disease genes was only about half that of AR disease genes ($P = 3 \times 10^{-4}$). We also found that, compared with mutations on AR genes, mutation pairs on different interaction interfaces of the same AD disease gene were much more likely to cause different disorders than mutation pairs on different interaction interfaces of the same AR disease gene (33% versus 19%, Fig. 1C). This result suggests an interrelationship between the interaction model and AD diseases, on account of different interactions of the same gene associated with different diseases. To increase the reliability of the results, we also analyzed the mutation data in M2SG database (Ji *et al.*, 2013), which includes curated mutations from OMIM and SwissProt. Among 162 genes having mutations on different interaction interfaces causing different diseases, 84 genes were AD genes, while 50 genes were AR genes ($P = 0.01$). As a comparison, among 168 genes having mutations on different interfaces causing the same disease, only 47 were AD genes, while 81 were AR genes ($P<0.01$). Mutation pairs on different interaction interfaces of the same AD gene were much more likely to cause different disorders compared with the mutation pairs on AR genes (17 versus 11%, $P<0.01$).

To investigate why some interacting genes cause different diseases, we asked whether tissue-specific expression imposes strong restriction for such interactions. Using global gene expression across 79 human tissues (see Section 2), we found that interacting genes causing the same disease share higher tissue homogeneity than interacting genes causing different diseases (Fig. 3E). We also found AD gene interactions and AR gene interactions for which both interacting genes have mutations on the same interaction interface and affecting the same disease have higher tissue homogeneity than interactions for which both interacting genes

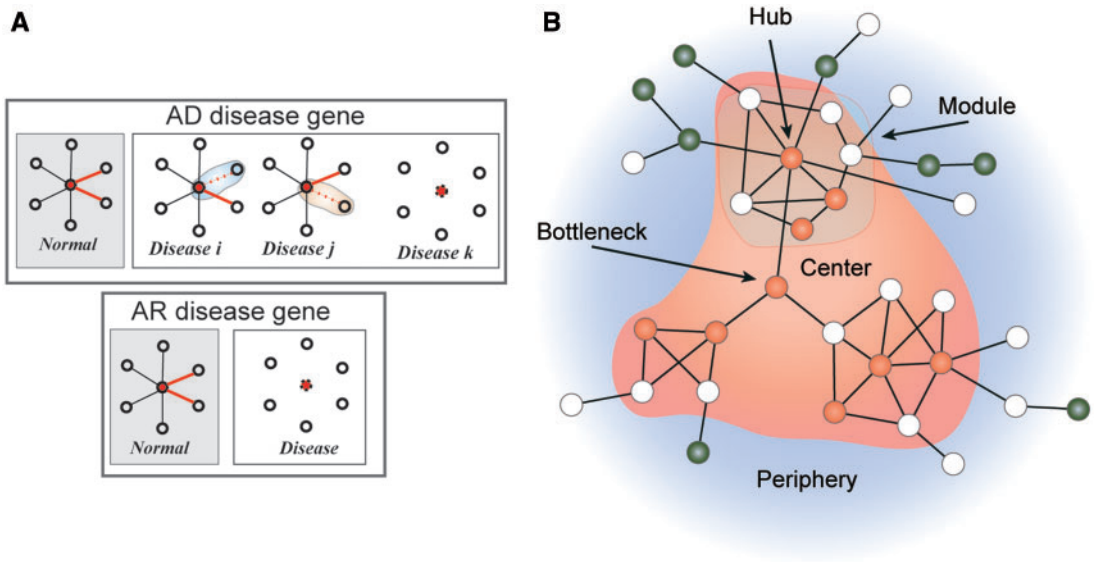

Fig. 4. Schematic illustration of AD/AR genes. (**A**) Schematic illustration of AD/AR genes with multiple interactions. Protein interaction can be used as a molecular basis to understand the correlation between network connectivity and AD disease, and the pleiotropy of AD disease genes, whereas it is not appropriate for explaining the pleiotropy for AR disease genes. (**B**) Schematic diagram of the differences between AD and AR disease genes. AD disease genes are shown as orange nodes, located in functional center of PIN as hubs, core members and bottlenecks of modules, whereas AR disease genes are shown as green nodes, segregated at the periphery of PIN, resulting in large network distances from other AR disease genes associated with different diseases

have mutations on the same interaction interface but affecting different diseases. As can be seen, tissue-specific expression imposes strong restriction for interactions. We also found that AR gene interactions have significantly higher tissue homogeneity than AD gene interactions.

In conclusion, the model suggests that protein interaction can be used as a molecular basis to understand the correlation between network connectivity and AD disease, and the pleiotropy of AD disease genes (Schematic illustration in Fig. 4A). This would give rise to a striking difference regarding the network position of AD versus AR genes (Fig. 4B). Generally, AD disease genes are located in functional center of PIN as hubs, core members and bottlenecks of modules, whereas AR disease genes are segregated at the periphery of PIN, resulting in large network distances from other AR disease genes associated with different diseases.

3.5 Network-based prediction of inheritance modes

Identifying inheritance modes traditionally depends on detailed pedigree information (Seton-Rogers, 2007). However, more and more *in silico* methods have been developed to predict new disease–gene associations (Ideker and Sharan, 2008), for which the pedigree data are usually unavailable. The topological features that differentiate AD diseases from AR diseases may help address this challenge. We thus evaluated such topological features and used SVM learning. A 5-fold cross-validation strategy was applied on those disease–gene associations with clear inheritance modes and with associated genes mapping into our PIN. The cross-validation process revealed a typical area under the curve of 0.85 and specificity, sensitivity, accuracy of 94, 63 and 84%, respectively, in predicting inheritance modes for disease–gene associations. This result can be used to assist the identification of inheritance modes. A typical example is sudden infant death (SID, OMIM: 272120). Evidence has been presented for the Mendelian association between SID and mutations on *SCN5A* (Weese-Mayer *et al.*, 2007). However, the inheritance mode of this association is unclear. We found that SCN5A is a hub gene and a bottleneck gene interacting with an AD gene ALB in our protein network, and also is a member of protein complex (HPRD: com_2971) that is significantly enriched by AD genes. We predicted the inheritance mode of SID-SCN5A association to be dominant. It has been known that mutations on SCN5A cause a series of heart diseases known as AD, including atrial fibrillation, brugada syndrome, cardiomyopathy, long QT syndrome and heart block.

4 DISCUSSION

Unlike non-pathological variations and out-frame disease mutations, it has been found recently that in-frame disease mutations are significantly enriched on interactions interfaces (Wang *et al.*, 2012). The in-frame mutations were suggested more likely to alter protein interaction and were found to be able to distinguish between AD and AR disorders (Zhong *et al.*, 2009). These suggest protein network can provide novel insights into genotype–phenotype correlations between different inheritance modes. We therefore analyzed a series of network properties between AD and AR inheritances and found (i) network connectivity strongly correlated with the likelihood of a gene causing AD diseases (disease risk) and the ability of a gene causing multiple AD diseases (pleiotropy); (ii) AD diseases are more likely to be associated with the bottleneck genes, whereas no significant

correlation was found for AR diseases. To explain the results, we refined a model based on previous studies and supported with more evidence.

It was argued that network-based study may improve our conceptual understanding of the term pleiotropy on account of diverse disorders that result from losing different interactions of the same gene (Chavali *et al.*, 2010; Tyler *et al.*, 2009). However, our study suggests that the pleiotropy of AR disease genes should be treated differently from AD disease genes. In addition to gene pleiotropy, the clinical features may show striking variation among patients for the same dominant disorder, and in some cases, show no abnormal clinical features (Cohn *et al.*, 2007; Milewicz *et al.*, 1998). We expect that, with further systematic investigations of the relationship between human AD disorders and gene networks, this variable expressivity and reduced penetrance of dominant diseases, together with the pleiotropy would be better understood. Pleiotropy is also one of the reasons that multiple diseases co-occur in the same patient. Many comorbid disease pairs have been shown to share the same pleiotropic gene (Park *et al.*, 2009). Therefore, our study may also suggest different comorbid tendencies between AD disease pairs and AR disease pairs.

Because topological central genes are more essential than other genes, our results may suggest that mutations on AD genes are more likely to cause severe impairment of cell function than mutations on AR genes. In our study, the proportion of AD genes having mouse lethal orthologs are significantly higher than that of AR genes (χ^2 test, $P<0.01$). The connection with important interaction may be one of the reasons why non-lethal dominant mutations are sometimes found to cause embryonic lethality when homozygous.

The results would be more consistent with the model by analyzing only loss-of-function mutations. For this purpose, we compiled 302 genes known to cause disease through haploinsufficiency (HI) from previous studies (Dang *et al.*, 2008; Seidman and Seidman, 2002). Similar results were reproduced by analyzing HI genes (Supplementary Fig. S6). Significantly different network properties were found between HI genes and AR genes, whereas no significantly different properties were found between HI genes and other AD genes.

A recent study has also investigated the relationship between dominant and recessive mutations in interfaces with the same or different disorders (Shao *et al.*, 2012). Different with this study, we only analyzed mutation pairs in the same gene because of the small number of interactions being mapped by mutations on the corresponding interfaces of interacting proteins. The study presented some different results and suggested that AD mutation pairs on corresponding interaction interfaces of interacting proteins are not likely to cause the same disease. We suggest that, it is possibly due to the fact that tissue-specific expression imposes spatial restriction on these interactions.

Gain-of-function mutations and dominant negative mutations are also important mechanisms of dominant diseases, for which the data are not currently available at a large scale. It is possible that some mutations on AD genes are associated with the gain of new interactions or that some dominant mutations give rise to altered gene product blocking the interaction of wild-type gene product. Revised model taking into account of multiple mutation types would be necessary in the future study. In addition, Disease progression is also associated with mutations causing the change of other molecular interaction types such as in transcriptional networks or metabolic networks. Integration of different types of molecular interactions should further improve our understanding of the inheritance modes of genetic diseases. We believe that, as more large-scale data of diverse molecular interactions and more specific disease mutations are detected, our knowledge of the genotype–phenotype relationship will be highly expanded, to which we hope our study made a contribution.

Funding: This work was supported by the Scientific Research Fund of Heilongjiang Provincial Education Department (NO:12541475).

Conflict of Interest: none declared.

REFERENCES

Barabasi,A.L. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.

Basu,M.K. *et al.* (2008) Evolution of protein domain promiscuity in eukaryotes. *Genome Res.*, **18**, 449–461.

Broeckel,U. and Schork,N.J. (2004) Identifying genes and genetic variation underlying human diseases and complex phenotypes via recombination mapping. *J. Physiol.*, **554**, 40–45.

Cattaneo,E. *et al.* (2001) Loss of normal huntingtin function: new developments in Huntington's disease research. *Trends Neurosci.*, **24**, 182–188.

Chatr-Aryamontri,A. *et al.* (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D816–D823.

Chavali,S. *et al.* (2010) Network properties of human disease genes with pleiotropic effects. *BMC Syst. Biol.*, **4**, 78.

Chiti,F. and Dobson,C.M. (2006) Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.*, **75**, 333–366.

Cohn,A.C. *et al.* (2007) Autosomal dominant optic atrophy: penetrance and expressivity in patients with OPA1 mutations. *Am. J. Ophthalmol.*, **143**, 656–662.

Dang,V.T. *et al.* (2008) Identification of human haploinsufficient genes and their genomic proximity to segmental duplications. *Eur. J. Hum. Genet.*, **16**, 1350–1357.

Das,J. *et al.* (2014) Exploring mechanisms of human disease through structurally resolved protein interactome networks. *Mol. Biosyst.*, **10**, 9–17.

Furlong,L.I. (2013) Human diseases through the lens of network biology. *Trends Genet.*, **29**, 150–159.

Giorgini,F. and Muchowski,P.J. (2005) Connecting the dots in Huntington's disease with protein interaction networks. *Genome Biol.*, **6**, 210.

Goh,K.I. *et al.* (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.

Hamosh,A. *et al.* (2005) Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.

Handel,A.E. *et al.* (2013) Next-generation sequencing in understanding complex neurological disease. *Expert Rev. Neurothe.r*, **13**, 215–227.

Ideker,T. and Sharan,R. (2008) Protein networks in disease. *Genome Res.*, **18**, 644–652.

Isserlin,R. *et al.* (2011) The Biomolecular Interaction Network Database in PSI-MI 2.5. *Database*, **2011**, baq037.

Ji,R. *et al.* (2013) M2SG: mapping human disease-related genetic variants to protein sequences and genomic loci. *Bioinformatics*, **29**, 2953–2954.

Kelley,R. and Ideker,T. (2005) Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.*, **23**, 561–566.

Kerrien,S. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.

Keshava Prasad,T.S. *et al.* (2009) Human protein reference database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

Ku,C.S. *et al.* (2011) Revisiting Mendelian disorders through exome sequencing. *Hum. Genet.*, **129**, 351–370.

Licata,L. *et al.* (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.*, **40**, D857–D861.

Margadant,C. and Sonnenberg,A. (2010) Integrin-TGF-beta crosstalk in fibrosis, cancer and wound healing. *EMBO Rep.*, **11**, 97–105.
Milewicz,D.M. *et al.* (1998) Reduced penetrance and variable expressivity of familial thoracic aortic aneurysms/dissections. *Am. J. Cardiol.*, **82**, 474–479.
Muers,M. (2010) Human disease: Edges, nodes and networks. *Nat. Rev. Genet.*, **11**, 4.
Park,J. *et al.* (2009) The impact of cellular networks on disease comorbidity. *Mol. Syst. Biol.*, **5**, 262.
Ross,E.D. *et al.* (2005) Prion domains: sequences, structures and interactions. *Nat. Cell. Biol.*, **7**, 1039–1044.
Rual,J.F. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.
Ruepp,A. *et al.* (2010) CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res.*, **38**, D497–D501.
Salwinski,L. *et al.* (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
Schuster-Bockler,B. and Bateman,A. (2008) Protein interactions in human genetic diseases. *Genome Biol.*, **9**, R9.
Seidman,J.G. and Seidman,C. (2002) Transcription factor haploinsufficiency: when half a loaf is not enough. *J. Clin. Invest*, **109**, 451–455.
Seton-Rogers,S. (2007) Patterns of inheritance. *Nat. Rev. Cancer*, **7**, 229–229.
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**, e31984.
Shy,M.E. *et al.* (2004) Phenotypic clustering in MPZ mutations. *Brain*, **127**, 371–384.
Stelzl,U. *et al.* (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.
Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
Taylor,I.W. *et al.* (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.*, **27**, 199–204.
Turner,B. *et al.* (2010) iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database (Oxford)*, **2010**, baq023.
Tyler,A.L. *et al.* (2009) Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *Bioessays*, **31**, 220–227.
Venkatesan,K. *et al.* (2009) An empirical framework for binary interactome mapping. *Nat. Methods*, **6**, 83–90.
Walhout,A.J. (2009) Getting an edge on human disease. *Mol. Syst. Biol.*, **5**, 322.
Walhout,A.J. *et al.* (2000) Protein interaction mapping in Celegans using proteins involved in vulval development. *Science*, **287**, 116–122.
Wang,X. *et al.* (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.*, **30**, 159–164.
Weese-Mayer,D.E. *et al.* (2007) Sudden Infant Death Syndrome: review of implicated genetic factors. *Am. J. Med. Genet. A*, **143A**, 771–788.
Xiao,S.J. *et al.* (2010) TiSGeD: a database for tissue-specific genes. *Bioinformatics*, **26**, 1273–1275.
Xu,Y. *et al.* (2011) 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.
Yu,H. *et al.* (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.*, **3**, e59.
Yu,H. *et al.* (2011) Next-generation sequencing to generate interactome datasets. *Nat. Methods*, **8**, 478–480.
Zhang,Y. *et al.* (2003) Depletion of wild-type huntingtin in mouse models of neurologic diseases. *J. Neurochem.*, **87**, 101–106.
Zhong,Q. *et al.* (2009) Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.*, **5**, 321.