European Journal of Human Genetics (2012) 20, 1162–1167
© 2012 Macmillan Publishers Limited All rights reserved 1018-4813/12
www.nature.com/ejhg

ARTICLE

Community of protein complexes impacts disease association

Qianghu Wang[1,2], Weisha Liu[1,2], Shangwei Ning[1], Jingrun Ye[1], Teng Huang[1], Yan Li[1], Peng Wang[1], Hongbo Shi[1] and Xia Li*,[1]

One important challenge in the post-genomic era is uncovering the relationships among distinct pathophenotypes by using molecular signatures. Given the complex functional interdependencies between cellular components, a disease is seldom the consequence of a defect in a single gene product, instead reflecting the perturbations of a group of closely related gene products that carry out specific functions together. Therefore, it is meaningful to explore how the community of protein complexes impacts disease associations. Here, by integrating a large amount of information from protein complexes and the cellular basis of diseases, we built a human disease network in which two diseases are linked if they share common disease-related protein complex. A systemic analysis revealed that linked disease pairs exhibit higher comorbidity than those that have no links, and that the stronger association two diseases have based on protein complexes, the higher comorbidity they are prone to display. Moreover, more connected diseases tend to be malignant, which have high prevalence. We provide novel disease associations that cannot be identified through previous analysis. These findings will potentially provide biologists and clinicians new insights into the etiology, classification and treatment of diseases.

European Journal of Human Genetics (2012) **20,** 1162–1167; doi:10.1038/ejhg.2012.74; published online 2 May 2012

Keywords: disease–disease association; protein complex; human disease network; disease comorbidity

INTRODUCTION

Molecular biology research has led to a great variety of knowledge about individual cellular components. It is increasingly evident that most cellular components carry out functions through intracellular and intercellular interactions with other cellular components.[1,2] This functional interconnectivity among molecular components implies that the impact of a specific disease-causing defect is not restricted to the activity of the cellular component that carries it, and can spread along the links of the interactome. It can also alter the activity of other cellular components, which in turn can cause other diseases. Hence, it is difficult to consider diseases as absolutely independent of others at the molecular level, and disease-causing defects may trigger cascades of failures that lead to the co-emergence of multiple diseases in a patient.

Furthermore, the emergence of a disease is rarely a consequence of an abnormality in one single cellular component, but rather reflects the interruptions of the complex intra- and intercellular network that connects tissue and organ systems.[3] For some diseases, disease-associated genes are functionally related to each other, in the form of protein complexes or biological pathways, and are consistent with the modular view of disease-associated genes.[3–6] For example, the genes *FANCA, FANCB, FANCC, FANCE, FANCF, FANCG, FANCL* and *FANCM*, which are all associated with Fanconi anemia, constitute the FA complex and carry out their functions in the form of this complex.[7,8] Hence, defects in different genes may result in similar disease phenotypes. Exploring disease associations could potentially open new opportunities for understanding the human diseasome and offer insights into new approaches to disease prevention, diagnosis and treatment.

Modern biology and medicine face a significant challenge in exploring the relationship between a disease phenotype and the underlying cellular perturbation.[6,9,10] Network-based approaches to human disease have been proposed as a platform from which to systematically explore the complexity of a particular disease at the molecular level and the molecular relationships among distinct disease phenotypes. For example, Goh *et al*[11] have created a human disease network (HDN) in which two diseases are connected if they share one or more disease-associated genes. Lee *et al*[12] have constructed a metabolic disease network in which two diseases are linked if the enzymes associated with them catalyze related metabolic reactions. Based on the observations above, we combined information about the cellular interactions, disease–gene associations and protein complex–gene associations to obtain statistically significant associations between diseases and generate a disease network in which two diseases are linked if there exists at least one protein complex that is associated with both diseases. Next, we tested the validity of the proposed associations between the diseases by exploring the degree to which the predicted disease relationships resulted in detectable disease comorbidity patterns in patients. The results indicate that the predicted disease associations can be frequently observed in patients, and disease pairs that are more interconnected in the disease network than others display higher comorbidity. Our findings not only potentially help us understand how different diseases are related based on their underlying molecular mechanisms but also provide

[1]College of Bioinformatics Science and Technology and Bio-pharmaceutical Key Laboratory of Heilongjiang Province, Harbin Medical University, Harbin, China
[2]The two authors contribute equally to this work, they are joint first authors.
*Correspondence: Professor X Li, College of Bioinformatics Science and Technology, Bio-pharmaceutical Key Laboratory of Heilongjiang, Harbin Medical University, Baojian Road No. 157 42#, Harbin 150081, China. Tel: +86 451 86615922; Fax: +86 451 86615922; E-mail: lixia@hrbmu.edu.cn

Received 31 October 2011; revised 6 March 2012; accepted 27 March 2012; published online 2 May 2012

insights into the design of novel, protein complex-guided therapeutic interventions for diseases.

MATERIALS AND METHODS

Disease–gene association data set

The disease–gene associations list was compiled from the Online Mendelian Inheritance in Man (OMIM)[13] database, which is a commonly used disease–gene association database.[11,12,14] As of May 2010, the list contained 5284 disease–gene associations, involving 1478 diseases and 3009 disease-associated genes (see Supplementary Table S1 for more detail).

Protein complex data set

A protein complex is a group of associated polypeptide chains in which proteins are connected by non-covalent protein–protein interactions. Generally, these protein complexes are functional units of many biological processes, and together they form all kinds of molecular machinery that carry out many biological functions. Protein complexes used in this study were experimentally verified human protein complexes compiled from the Comprehensive Resource of Mammalian protein complexes database (CORUM)[15] at the Munich Information Center for Protein Sequences.[16] The CORUM database offers a commonly used resource of manually annotated protein complexes from mammalian organisms. In all, we obtained 1343 human protein complexes in the core data set involving a total of 2312 unique genes (Supplementary Table S1).

Disease comorbidities data set

To test the validity of the proposed disease associations, we examined the disease pairs that our analysis found to be linked using disease co-occurrence information at the population level. We obtained statistically significant pairwise comorbidity associations reconstructed from over 30 million medical records in the US Medicare claims database, which are frequently used for epidemiological and demographic studies[17,18] and which contain information on 13 039 018 elderly patients. The comorbidity strength was quantified using two measures: the relative risk (RR) and the phi-correlation (ϕ) (see Supplemental Text for more detail). This data set has been reported by Hidalgo *et al.*[19] To reduce the impact of the extreme elements within the comorbidity data source, we filtered the original comorbidity association data according to the data distribution (see Supplemental Text for more detail). For our purposes, we selected comorbidity associations with $RR < 100$ and $|\phi| < 0.05$ for further analysis.

Establish disease–protein complex association based on constituent genes

To construct the human disease–disease associations, we first established disease–protein complex associations using the disease–gene associations in the OMIM database, given that the data relating complete protein complex to diseases is quite limited. We examined the overlap between a disease and a protein complex by looking at the constituent genes and established the links between the diseases and the protein complexes. We used a one-sided Fisher's exact test to evaluate the overlap between a disease and a protein complex in terms of their constituent genes. Afterwards, the raw P-values were adjusted by using the Benjamini–Hochberg procedure to control the false discovery rate.[20] We selected disease–protein complex pairs with adjusted P-values < 0.05 as significantly associated pairs for further analysis.

In addition, we evaluated the biological diversity of protein complexes that were associated with the same disease and rank the diseases based on their associated protein complex content index (PCCI)[21] (see Supplemental Text for more details). Similarly, we also evaluated the biological diversity of diseases that were associated with the same protein complex and ranked the protein complexes based on their associated disease content index (DCI)[21] (see Supplemental Text).

Establish disease–disease association based on the community of protein complexes

Next, we used the obtained list of significantly associated disease–protein complex pairs to identify associations between diseases. Close physical interactions, expression correlations and functional communications could occur among the protein subunits of a protein complex.[22–24] Thus, the mutation of one protein subunit may be propagated to other protein subunits within the same protein complex. In other words, a malfunction in a protein complex may yield the dysfunction of multiple protein subunits. Therefore, multiple diseases may be caused by the malfunction of a protein complex. Based on this biological mechanism, we used the obtained list of significantly associated disease–protein complex pairs in the first step to identify correlations between diseases. We inferred that two diseases are potentially related to each other if they share one or more commonly associated protein complexes (one example we used to illustrate the hypothesis of our study is depicted in Supplemental Text).

When two diseases share more than one protein complex, there may be redundant content among these shared protein complexes. Taking this redundancy into account, we measured the strength of the association between two diseases, d_1 and d_2, as follows:

$$S_{d_1,d_2} = \frac{|\mathrm{SPC}|}{\frac{G(\mathrm{SPC})}{U(\mathrm{SPC})}},$$

where $\mathrm{SPC} = \{\mathrm{PC}_1, \mathrm{PC}_2 \ \ldots \ \mathrm{PC}_n\}$ denotes the set of protein complexes shared by d_1 and d_2, $|\mathrm{SPC}|$ indicates the size of SPC, $G(\mathrm{SPC}) = \sum_{i=1}^{n} |\mathrm{PC}_i|$, $U(\mathrm{SPC}) = \left|\bigcup_{i=1}^{n} \mathrm{PC}_i\right|$, and $|\mathrm{PC}_i|$ denotes the number of genes involved in PC_i. S_{d_1,d_2} is equal to 1 if the protein complexes in set SPC are completely redundant, while $S_{d_1,d_2}$is equal to the size of set SPC if there is no redundant content among these shared protein complexes.

To identify potentially novel disease relationships that cannot be identified in previous studies, we compared our results to those gained through the shared gene hypothesis and the shared pathway hypothesis (pathway data obtained from the KEGG database and the Biocarta database) on the same disease data input.

Comorbidity analysis

Disease pathogenesis results from the breakdown of physiological cellular processes, including interactions among components of the genome, proteome, metabolome and environment. Hence, the activities of the affected protein complexes are likely to contribute to disease progression and comorbidity at the molecular level. To examine whether disease pairs that are related to each other based on their associated protein complex(es) have comorbid tendencies, we analyzed the relationship between disease association strength and disease co-occurrence.

It is noteworthy that the disease names in the list of comorbidity associations between pairwise diseases are identified by ICD-9-CM codes. Therefore, we also mapped the OMIM disease names into ICD-9-CM codes, which is consistent with previous studies.[12,14,19] For a detailed list of the ICD-9 codes, http://www.icd9data.com. Although some diseases, such as β-ureidopropionase deficiency and Fechtner syndrome, cannot be assigned an ICD-9-CM code, 1090 out of the 1478 diseases extracted from the OMIM database do map reliably to ICD-9-CM codes. We drew the available comorbidity correlations corresponding to the connected disease pairs that were indicated by our results.

Here, we used both comorbidity measures to ensure the robustness of our results. To quantify the degree of comorbidity caused by the observed disease associations, we measured the average RR and ϕ for disease pairs that our method indicated were linked at the molecular level. In addition, we compared the comorbidity tendencies of diseases linked by shared protein complexes to those linked by shared genes and shared pathways (see Supplemental Text).

RESULTS

Associations between diseases and protein complexes

In total, 763 protein complexes were mapped to 447 diseases, and 2238 disease–protein complex associations were generated. We calculated the distribution of the diseases according to the number of associated protein complexes (Supplementary Figure S1A) and the distribution of protein complexes according to the number of associated diseases (Supplementary Figure S1B). On average, a disease was linked to about five protein complexes (median = 2), and a protein complex was linked to about three diseases (median = 2).

As shown in Supplementary Table S2, we ranked the diseases according to their associated PCCI. Diseases that are connected to many protein complexes and therefore easily caused by many biological processes are at the top of this list, including malignant diseases with high prevalence, such as colorectal cancer, breast cancer, pancreatic cancer, hepatocellular cancer and leukemia.[25] On the other hand, some diseases are linked to only a few protein complexes and can be caused by defects in few specific biological processes, such as adrenoleukodystrophy, Caffey disease and Dysautonomia, are all associated with only one single protein complex.

Similarly, we evaluated the diversity of the diseases associated with the same protein complex. We also ranked the protein complexes based on their associated DCI (see Supplementary Table S2). The top of the list is primarily composed of signal transduction protein complexes and various protein complexes involved in enzymatic activity regulation, the immune response, metabolic processes and structural complexes that often form the molecular machinery and are involved in many different biological processes. On the basis of this analysis, it was clear that a protein complex can be connected to a set of very different diseases, which would indicate that the diverse diseases likely had a common biological mechanism. For example, the Er–α–p53–hdm2 complex is linked to many different diseases, such as glycolipid metabolic diseases and multiple types of cancers. Moreover, many diseases from these two disease classes are associated with each other through this protein complex.

Associations between diseases

As a result, we captured 1953 associated disease-disease pairs, covering 404 diseases (Supplementary Table S3). For each associated disease pair, on average, two diseases are linked through two commonly associated protein complexes (median = 1, Supplementary Figure S1C).

Among all 1953 disease relationships, part of them could be discovered in previous studies, as they share one or more common genes or metabolic pathways. The rest are potentially novel disease relationships, as they can only be connected based on shared protein complexes. Table 1 presents some examples of potentially novel disease relationships (see Supplementary Table S3 for complete descriptions).

Significant comorbidity between the linked diseases

We obtained a total number of comorbidities for 608 linked disease pairs considered in our study (Supplementary Table S4). Compared with the disease pairs that do not share protein complexes, we found nearly a twofold increase in the average comorbidity of the disease pairs that share protein complexes (Figure 1a). This suggests that, if a patient develops a particular disease associated with at least one protein complex, he or she has about a twofold higher chance of developing other diseases that share common associated protein complex(es) relative to diseases that do not.

Next, we discussed whether disease pairs that are more interconnected in the HDN show higher comorbidity. To clarify this, in Figure 1b we show that, on average, comorbidity increases rapidly with the strength of disease association. This observation indicates that the disease pairs that are related to each other based on shared protein complexes have a comorbid tendency. The observed associations between diseases may help us to identify new comorbidity patterns along their potential genetic origin. After examining the entire set of 1953 disease pairs that are genetically linked in our results, we found some disease pairs whose comorbidity patterns are already well known to the medical community, for example, hypertension and ischemic stroke[26,27] or diabetes mellitus and chronic anemia.[28] These examples demonstrate that the capacity of the protein complex-based approach in identifying potentially interesting disease pairs is worth of further research.

In addition, we also examined whether the protein complex-based method is indeed a valid method of discovering novel disease relationships by analyzing the potential comorbidity of the diseases that were linked in our study. For example, we found some novel disease associations, Muir–Torre syndrome and Xeroderma pigmentosum ($RR = 11.868$, $\phi = 0.001$), and Rhabdoid tumors and Walker–Warburg syndrome ($RR = 15.651$, $\phi = 0.013$) have a strong comorbidity effect in the human population by combining information on population-level disease comorbidity patterns extracted from Medicare data. These disease pairs can be linked based on shared protein complexes, but these associations would not have been found using the shared gene-based or shared pathway-based approaches. These examples support our hypothesis that protein complexes can be used to predict disease associations and determine novel disease associations that other methods cannot capture.

Construction of the protein complex-based HDN

Here, we generated a protein complex-based HDN (PC-HDN) in which two diseases are linked if they share at least one associated protein complex, and this can be seen as a map summarizing associations between diseases. The PC-HDN consists of 404 disease nodes and 1953 edges (disease relationships). Figure 2 shows a filtered version of the PC-HDN in which 711 disease relationships with the number of shared protein complexes > 1 were displayed.

Table 1 Examples of novel disease relationships

Disease 1	*Disease 2*	*Shared disease-related protein complex(es)*
Walker–Warburg syndrome	Rhabdoid tumors	Emerin complex 32
Muir–Torre syndrome	Xeroderma pigmentosum	ERCC1–ERCC4–MSH2 complex
α-1-Antichymotrypsin deficiency	Histiocytoma, angiomatoid fibrous	ACT–CREB complex
Alzheimer disease	Huntington disease	PRNP–apolopoproteinE3 complex
Angiofibroma, somatic	Thrombophilia	MLL–HCF complex
Bloom syndrome	Tyrosinemia	BRAFT complex FA complex
Estrogen resistance	Adrenal cortical carcinoma	Er–α–p53–hdm2 complex
Charcot–Marie–Tooth disease	Optic atrophy	Mediator complex
van der Woude syndrome	Aortic valve disease	HES1 promoter–Notch enhancer complex
Parathyroid adenoma	Refsum disease	Paf complex

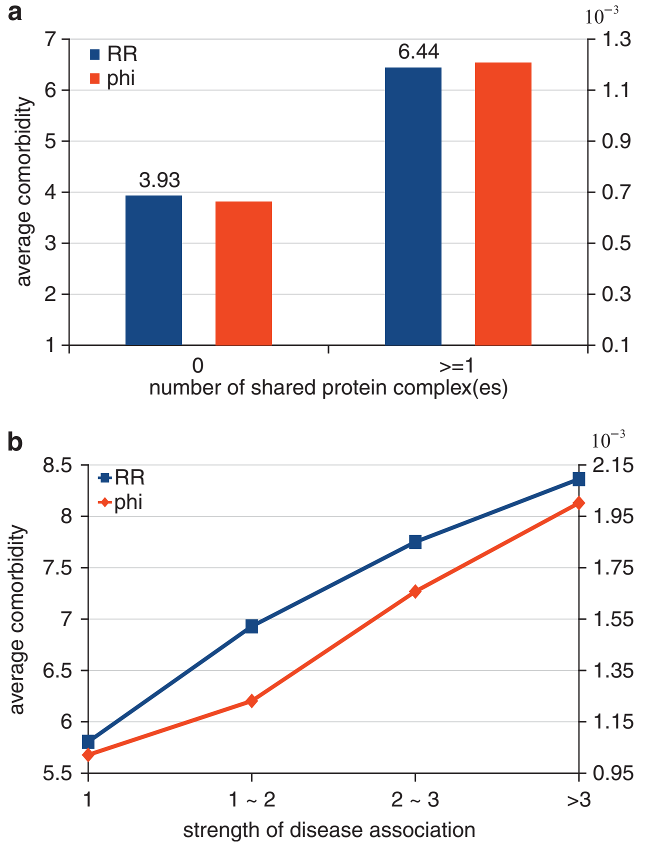

Figure 1 The relationship between linked disease pairs and their comorbidities. (**a**) Comparison of average comorbidity for pairwise diseases that share at least one protein complex and for pairwise diseases that share no protein complex in our study. (*RR*, blue; ϕ, red). (**b**) Average comorbidity for linked disease pairs with the increasing strength of their associations, which are quantified by S_{d_1,d_2}.

Under our hypothesis, this network not only could capture the disease–disease associations that have been discovered in previous research but could also reveal potentially novel disease–disease associations that are only based on protein complexes. Some interesting examples of disease relationship clusters in this PC-HDN are displayed in detail. One example reveals a central theme of the relationships between abnormality in glucolipid metabolism and cancers, in which atherosclerosis, coronary artery disease, diabetes mellitus, hypertension, myocardial infarction and many kinds of cancers are involved (Figure 3a). Myocardial infarction and nasopharyngeal carcinoma are offered as detailed examples (Figure 3b). They are linked to each other by sharing one protein complex, Er–α–p53–hdm2 complex. This protein complex participates in enzymatic activity regulation. In fact, increasing evidence indicates that glucolipid metabolic diseases are closely related to carcinogenesis and cancer development.[29,30] In addition, Figure 3c shows another disease cluster that consists of immunodeficiency-related diseases and multiple types of cancers. Previous studies have demonstrated that somatic immunodeficiency is associated with human cancer.[31,32] In our study, we found that some immunodeficiency-related diseases, such as autoimmune lymphoproliferative syndrome, HIV infection and so on, are connected to many diseases through common protein complexes. Autoimmune lymphoproliferative syndrome and leukemia are treated as an example and are presented in Figure 3d. They are connected based on sharing the BAR–BCL2–CASP8 complex. This protein complex involves in induction of proapoptotic gene products. Taken together, these findings open new opportunities in biomolecular and bioinformatics approaches to diseases.

Characterizing the PC-HDN

This PC-HDN is a densely connected network with an average clustering coefficient of 0.701 and an average shortest path length between any two diseases of 3.45. The low average shortest path length and the large clustering coefficient indicated that the PC-HDN had the small-world properties of most biologically complex networks.[3] The network was scale-free, as the degree distribution had a power-law tail. We ranked the diseases based on their degrees in the PC-HDN. As shown in Supplementary Table S5, the top of the list consists of a diverse array of diseases from the OMIM database that are connected to many other diseases, and the great majority of them are malignant diseases that have high lethality, such as cancer, myocardial infarction and so on. This observation suggests that more-connected diseases are prone to be more lethal, and when patients develop highly connected diseases they are likely to be at an advanced stage of disease, as the connected diseases can be reached through multiple paths in the PC-HDN.

We manually classified the diseases into 22 classes based on the physiological systems affected, according to the classification reported by Goh *et al.*[11] As the disease data in OMIM have been updated, each class contains more diseases than were reported by Goh *et al*. In the PC-HDN, there were 503 associations between diseases within the same disease class, which is, on average, a threefold enrichment compared with the 156 links (empirical P-values $< 1e-04$) found in a random disease network (the disease node labels were randomised, this randomization was performed 10 000 times, and the empirical P-values was the probability of obtaining more associations between diseases within the same disease class in the randomised networks than in the actual PC-HDN) (Supplementary Figure S2). This result indicates that within this PC-HDN, diseases from the same disease class tend to relate to each other at the global level.

It is known that if two diseases have associated comorbidity, the occurrence of one of them in a patient may increase the likelihood of developing the other disease.[3,12,14,19] Disease comorbidity research shows that the disease pairs that were linked in our study have comorbid tendencies; using this result, the map allows us to explore disease progression as a network process in which patients tend to develop diseases that are close (according to the PC-HDN links) to those they already have.

DISCUSSION

Biomedical researchers have focused on the commonality of the pathology or etiology of diseases over the past decade. Several resources have been established to help reveal relationships among diseases. The combination of molecular biology, genetics and clinical medicine has greatly facilitated understanding of how different diseases relate to each other. Enormous efforts have been devoted to molecular-based methods for studying disease associations at the molecular level. Although these methods are tremendously successful, they are far from sufficient and complete, and it is still a challenge to identify relationships between diseases.

Here, we proposed a novel approach for exploring the relationships between human diseases based their associations with protein complexes. Based on shared protein complexes, our disease network can disclose potentially novel disease relationships that have not been captured by previous methods. There are also some disease relationships that can be found through other methods but the protein

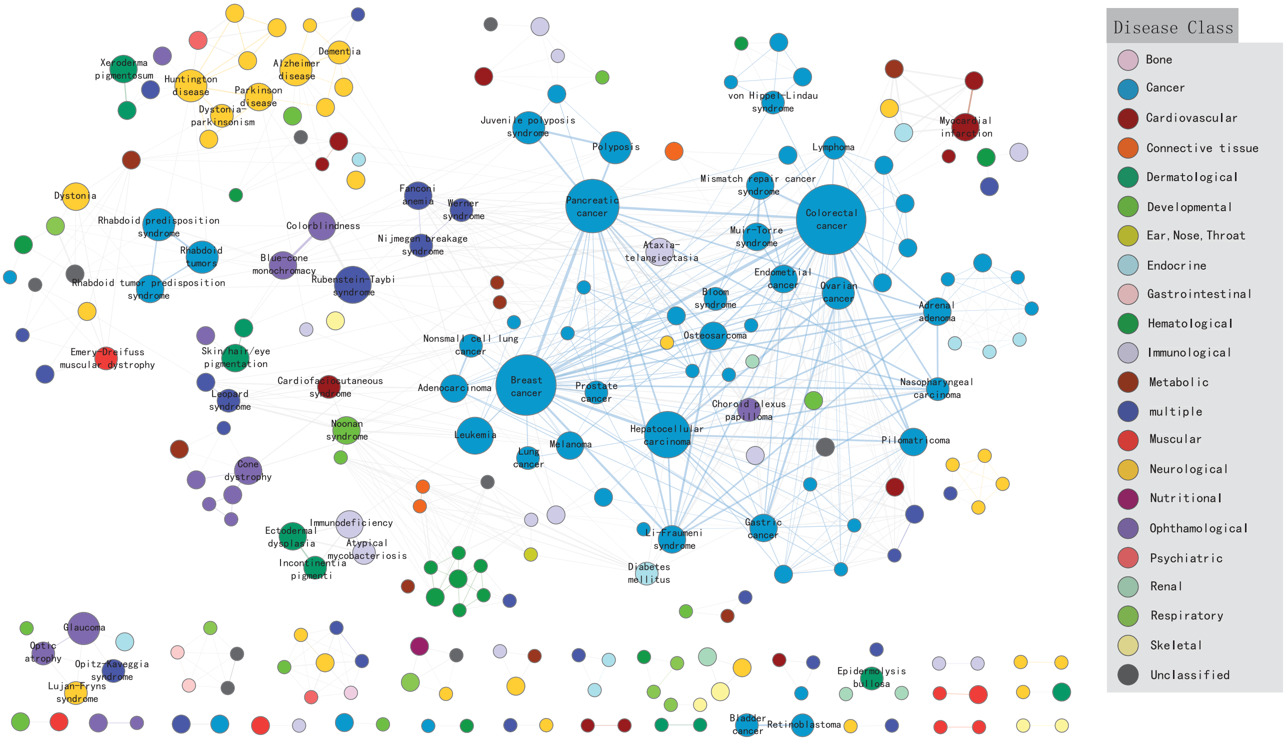

Figure 2 A filtered PC-HDN. Each node is colored based on the disease class to which it belongs, and the size of it is proportional to the number of protein complexes associated with the corresponding disease. Edges linking diseases within the same disease class are colored with a dimmer color of the corresponding disease class, and edges linking different disease classes are light gray. The edge thickness is proportional to the disease association strength measured by S_{d_1,d_2}. The names of diseases with ≥10 associated protein complexes are displayed.

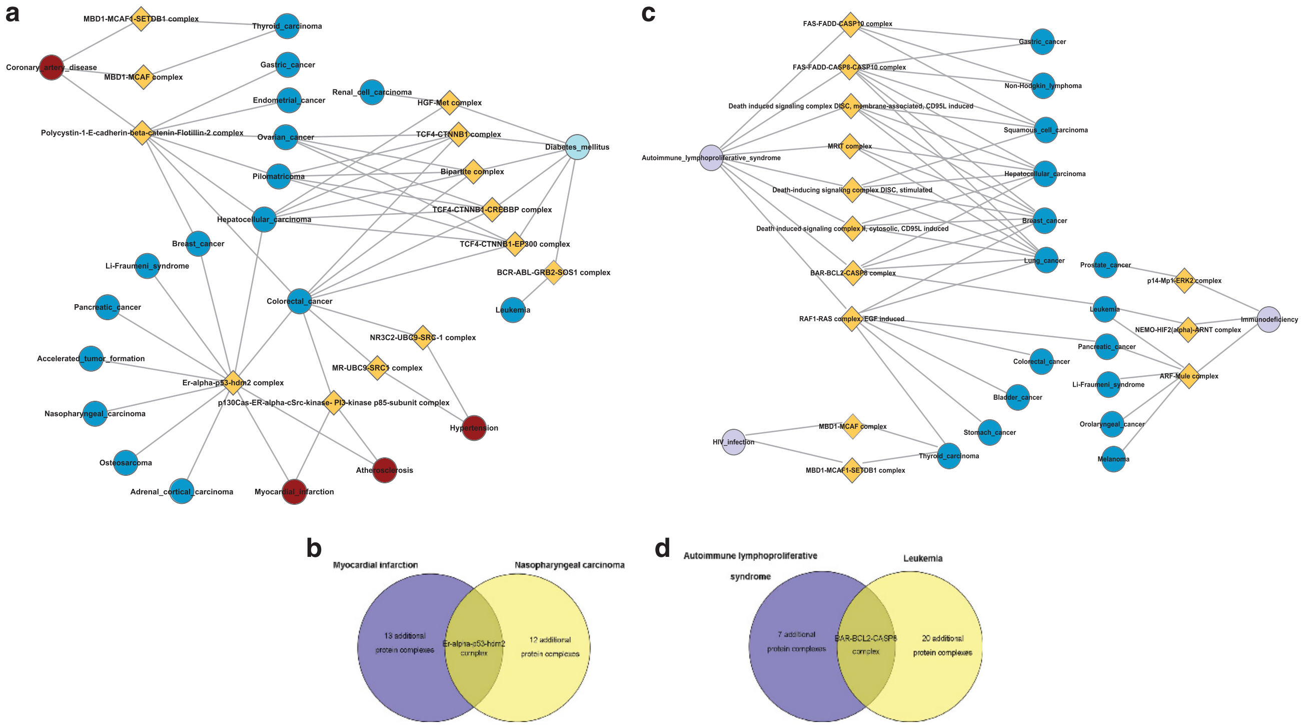

Figure 3 Two examples of disease clusters derived from the PC-HDN. (**a**) A disease cluster consisting of five glucolipid metabolism-related diseases plus many kinds of cancers, and protein complexes (orange diamond) linking two diseases from these two disease categories. (**b**) Myocardial infarction and nasopharyngeal carcinoma offer a detailed example in Figure 3b. (**c**) Another disease cluster consists of three immunodeficiency-related diseases and multiple types of cancers, and protein complexes linking two diseases from these two disease categories. (**d**) Autoimmune lymphoproliferative syndrome and leukemia offer a detailed example in Figure 3**d**.

complex approach missed. The latter disparity could result from a lack of protein complex data, but the amount of available protein complex data is ever increasing and the protein complex-based research will continue to provide novel and informative disease associations for use by biologists and clinicians. However, in its current state, the PC-HDN can still capture potential candidates for novel disease relationships that are complementary to those obtained using other approaches. These novel disease relationships captured by PC-HDN not only fill the gaps in our theoretical and experimental knowledge of diseases but also offer new insights into disease etiology, classification and associated gene identification.

Using networks to display disease relationships have multiple potential biological and clinical applications. Our results show that disease progress can be represented along the links in the HDN by using this network method. Studying the structure of the HDN may help us to predict disease outcomes and to identify tailored therapeutic strategies. Moreover, this protein complex-based approach to diseases can aid in drug discovery, particularly if one drug is already approved to treat a disease through regulating the activity of a protein complex and, therefore, can potentially be used to treat other diseases that are linked to the same protein complex.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This research was supported by the National Natural Science Foundation of China (grant nos. 31100948, 30871394 and 91029717), the National Science Foundation of Heilongjiang Province (grant nos. D201114, QC2009C23), the Science Foundation of Educational Commission of Heilongjiang Province (grant no. 11551233), the Graduate Innovation Fund of Heilongjiang Province (grant nos. YJSCX2011-334HLJ and YJSCX2009-226HLJ).

1 Barabasi AL, Oltvai ZN: Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004; **5**: 101–113.
2 Friedman A, Perrimon N: Genetic screening for signal transduction in the era of network biology. *Cell* 2007; **128**: 225–231.
3 Barabasi AL, Gulbahce N, Loscalzo J: Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011; **12**: 56–68.
4 Oti M, Brunner HG: The modular nature of genetic diseases. *Clin Genet* 2007; **71**: 1–11.
5 Barabasi AL: Network medicine from obesity to the 'diseasome'. *N Engl J Med* 2007; **357**: 404–407.
6 Loscalzo J, Kohane I, Barabasi AL: Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Mol Syst Biol* 2007; **3**: 124.
7 de Winter JP, van der Weel L, de Groot J *et al*: The Fanconi anemia protein FANCF forms a nuclear complex with FANCA, FANCC and FANCG. *Hum Mol Genet* 2000; **9**: 2665–2674.
8 D'Andrea AD: The Fanconi anemia/BRCA signaling pathway: disruption in cisplatin-sensitive ovarian cancers. *Cell Cycle* 2003; **2**: 290–292.
9 Lamb J, Crawford ED, Peck D *et al*: The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006; **313**: 1929–1935.
10 Lage K, Karlberg EO, Storling ZM *et al*: A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 2007; **25**: 309–316.
11 Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL: The human disease network. *Proc Natl Acad Sci USA* 2007; **104**: 8685–8690.
12 Lee DS, Park J, Kay KA, Christakis NA, Oltvai ZN, Barabasi AL: The implications of human metabolic network topology for disease comorbidity. *Proc Natl Acad Sci USA* 2008; **105**: 9880–9885.
13 Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005; **33**: D514–D517.
14 Park J, Lee DS, Christakis NA, Barabasi AL: The impact of cellular networks on disease comorbidity. *Mol Syst Biol* 2009; **5**: 262.
15 Ruepp A, Waegele B, Lechner M *et al*: CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res* 2009; **38**: D497–D501.
16 Mewes HW, Dietmann S, Frishman D *et al*: MIPS: analysis and annotation of genome information in 2007. *Nucleic Acids Res* 2008; **36**: D196–D201.
17 Lauderdale DS, Furner SE, Miles TP, Goldberg J: Epidemiologic uses of Medicare data. *Epidemiol Rev* 1993; **15**: 319–327.
18 Mitchell JB, Bubolz T, Paul JE *et al*: Using Medicare claims for outcomes research. *Med Care* 1994; **32**: JS38–JS51.
19 Hidalgo CA, Blumm N, Barabasi AL, Christakis NA: A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol* 2009; **5**: e1000353.
20 Thissen D, Steinberg L, Kuang D: Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *J Educ Behav Stat* 2002; **27**: 77–83.
21 Li Y, Agarwal P: A pathway-based view of human diseases and disease relationships. *PLoS One* 2009; **4**: e4346.
22 Szilagyi A, Grimm V, Arakaki AK, Skolnick J: Prediction of physical protein-protein interactions. *Phys Biol* 2005; **2**: S1–16.
23 Cai L, Xue H, Lu H, Zhao Y, Zhu X, Bu D, Ling L, Chen R: Analysis of correlations between protein complex and protein-protein interaction and mRNA expression. *Chinese Sci Bull* 2003; **48**: 2226–2230.
24 Yunku Yeu J, Youngmi Y, Sanghyun P: Protein complex discovery from protein interaction network with high false-positive rate. *Lect Notes Comput Sci* 2011; **6623**: 177–182.
25 American Cancer Society 2011: *Cancer Facts and Figures 2011,* 2011.
26 Wallace JD, Levy LL: Blood pressure after stroke. *JAMA* 1981; **246**: 2177–2180.
27 Hachinski V: Hypertension in acute ischemic strokes. *Arch Neurol* 1985; **42**: 1002.
28 Thomas MC, MacIsaac RJ, Tsalamandris C, Power D, Jerums G: Unrecognized anemia in patients with diabetes: a cross-sectional survey. *Diabetes Care* 2003; **26**: 1164–1169.
29 Mulligan HD, Beck SA, Tisdale MJ: Lipid metabolism in cancer cachexia. *Br J Cancer* 1992; **66**: 57–61.
30 Shaw RJ: Glucose metabolism and cancer. *Curr Opin Cell Biol* 2006; **18**: 598–608.
31 Cottier H, Hess MW, Walti ER: Immunodeficiency and cancer: mechanisms involved. *Schweiz Med Wochenschr* 1986; **116**: 1119–1126.
32 Hadden JW: Immunodeficiency and cancer: prospects for correction. *Int Immunopharmacol* 2003; **3**: 1061–1071.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)