

ARTICLE

Prioritising risk pathways of complex human diseases based on functional profiling

Yan Li^{1,4}, Teng Huang^{1,4}, Yun Xiao¹, Shangwei Ning¹, Peng Wang¹, Qianghu Wang¹, Xin Chen¹, Xu Chaohan¹, Donglin Sun², Xia Li^{*,1} and Yixue Li^{*,3}

Analysis of the biological pathways involved in complex human diseases is an important step in elucidating the pathogenesis and mechanism of diseases. Most pathway analysis approaches identify disease-related biological pathways using overlapping genes between pathways and diseases. However, these approaches ignore the functional biological association between pathways and diseases. In this paper, we designed a novel computational framework for prioritising disease-risk pathways based on functional profiling. The disease gene set and biological pathways were translated into functional profiles in the context of GO annotations. We then implemented a semantic similarity measurement for calculating the concordance score between a functional profile of disease genes and a functional profile of pathways (FPP); the concordance score was then used to prioritise and infer disease-risk pathways. A freely accessible web toolkit, 'Functional Profiling-based Pathway Prioritisation' (FPPP), was developed (<http://bioinfo.hrbmu.edu.cn/FPPP>). During validation, our method successfully identified known disease–pathway pairs with area under the ROC curve (AUC) values of 96.73 and 95.02% in tests using both pathway randomisation and disease randomisation. A robustness analysis showed that FPPP is reliable even when using data containing noise. A case study based on a dilated cardiomyopathy data set indicated that the high-ranking pathways from FPPP are well known to be linked with this disease. Furthermore, we predicted the risk pathways of 413 diseases by using FPPP to build a disease similarity landscape that systematically reveals the global modular organisation of disease associations.

European Journal of Human Genetics (2013) 21, 666–672; doi:10.1038/ejhg.2012.218; published online 10 October 2012

Keywords: complex human diseases; prioritising risk pathways; functional profiling; concordance score

INTRODUCTION

Complex diseases result from the combined deregulation of many interacting genes rather than the mutation of an individual gene.^{1–4} The interactions and reactions of several genes constitute a biological pathway. The activation and inhibition of biological pathways directly affect the occurrence and development of complex diseases. For example, inhibition of the Notch pathway with a gamma-secretase inhibitor may provide a therapeutic benefit to a subset of ERBB2-positive breast cancer patients.⁵ Therefore, the identification of disease-associated pathways is very important.

Most existing pathway analysis approaches make use of the pathway annotation of genes to identify the overrepresentation of biological pathways. Using Fisher's exact test or the cumulative hyper-geometric test, pathways with significant overrepresentation are considered disease-related.⁶ The use of these tools, such as DAVID, GOEAST and PathwayExplorer, has become common.^{7–12} These pathway overrepresentation analysis approaches mainly depend upon mapping overlapping genes between pathways and diseases. However, some nonoverlapping genes may be functionally related between pathways and diseases. These approaches only utilise overlapping genes and ignore the functional association of nonoverlapping genes with pathways and diseases, thereby affecting the identification of

disease-risk pathways. We are interested in studying the global function of all genes in a set to identify disease-risk pathways.

On the basis of the above notion, we constructed a novel functional profiling-based method for identifying disease-risk pathways that considers the functional association between all genes of diseases and pathways. Functional profiling-based pathway prioritisation (FPPP) is a gene set analysis approach. First, a statistical analysis is conducted to identify significant GO terms and to construct the functional profile of the disease gene set (FPD). A biological pathway could also be translated into a functional profile of pathways (FPP). FPPP globally considers the functions of all genes and their statistical significance. Second, we implement a semantic similarity measurement, called a concordance score, to quantify the association between an FPD and an FPP that were created using GO. The concordance score indicates the functional association of the disease and the pathway. Finally, for each disease, we calculate the concordance scores for each biological pathway in a collection, and we rank the biological pathways. Plausible disease pathways are selected based on the scores. The detailed algorithms of FPPP are given in the 'Materials and Methods' section and in Figure 1.

To validate the performance of FPPP, we used several strategies that are discussed below. We demonstrated that our functional profiling-

¹College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China; ²Laboratory of Medical Genetics, Harbin Medical University, Harbin, China; ³Bioinformatics Center, Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

*Correspondence: Professor X Li, College of Bioinformatics Science and Technology, Harbin Medical University, 194 Xuefu Road, Harbin 150081, China. Tel: 86 451 86615922; Fax: 86 451 86615922; E-mail: lixia@hrbmu.edu.cn

or Dr Y Li, Bioinformatics Center, Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, China. Tel: 86 21 54065001; Fax: 86 21 54065001; E-mail: yxli@sibs.ac.cn

⁴These authors contributed equally to this work.

Received 10 April 2012; revised 7 August 2012; accepted 28 August 2012; published online 10 October 2012

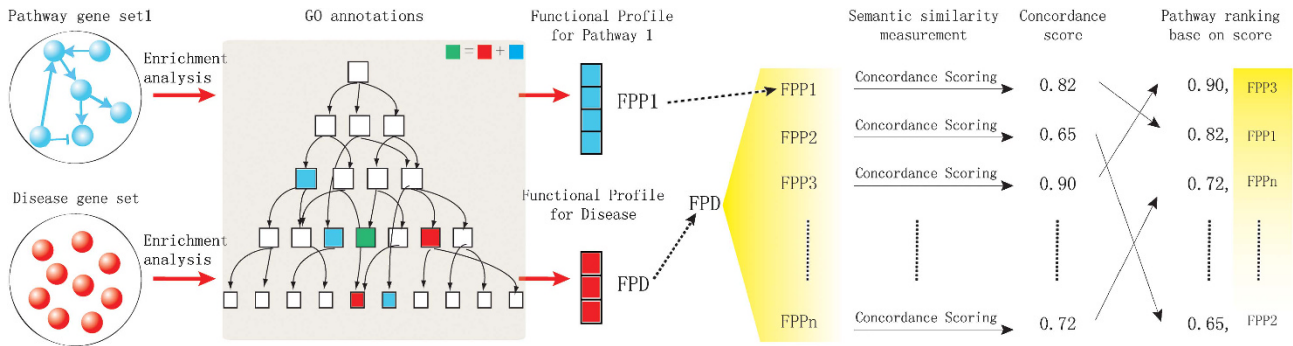


Figure 1 A workflow diagram of the FPPP method. In the first step, the disease gene set and biological pathway 1 were translated into an FPD and an FPP1 using the functional profiling translation. Then, to quantify the association between an FPD and an FPP1, the semantic similarity measurement was conducted to calculate their concordance score. Finally, for each disease, the concordance scores of 186 biological pathways were calculated and prioritised to infer risk pathways. A workflow diagram of an example of how the concordance score between DCM genes and DCM pathway is calculated is shown in supplementary Figure S1.

based approach is reliable and efficient for prioritising disease-risk pathways using known disease–pathway associations. In pathway prioritisations of 18 diseases, 13 diseases had the corresponding true pathways ranked within the top 20, and the common and specific-risk pathways among these diseases were identified. A case study of dilated cardiomyopathy (DCM) indicated that the high-ranking pathways from FPPP were well documented as being relevant to this disease. Moreover, we show that the risk pathways for 413 diseases inferred using FPPP revealed the landscape of disease similarity on a biological basis. In addition, a web toolkit, named ‘FPPP’, was developed based on this method (<http://bioinfo.hrbmu.edu.cn/FPPP>). Users of this toolkit submit a disease gene set or a name of a disease and receive a global prioritisation of biological pathways. The computational framework described here offers an alternative method for disease pathway discovery to aid further experimental research.

MATERIALS AND METHODS

Pathway data and disease genes

Biological pathways were downloaded from the GSEA website (<http://www.broadinstitute.org/gsea/>),¹³ including those from KEGG,¹⁴ Biocarta and Reactome.¹⁵ The web toolkit provides the pathway prioritisation for three databases. A pathway is composed of several molecular interactions and reactions and is used as a gene set in our method.

Disease genes were downloaded from the genetic association database (GAD), which contains all known gene–phenotype associations and includes common complex non-Mendelian diseases.¹⁶ We obtained 413 disease-related gene sets by trimming and merging all gene–phenotype relationships. We can also obtain disease classes from GAD. Each disease falls into a disease class, and the 413 total diseases fall into 18 major disease classes.¹⁷

Thirty-six known human disease pathways from KEGG provide an ideal benchmarking data set for pathway analysis. We obtained known associations of 26 pathways corresponding to 18 disease gene sets by screening and merging common diseases related to pathways in KEGG and related to phenotypes in GAD. Each of the known disease–pathway associations is taken as one test case, and we calculate the similarity score for each case (see Supplementary Table S1). For benchmark tests, the known disease pathway was treated as unknown in a random pathway list to compute an empirical permutation *P*-value.

Obtaining the FPD and FPP based on GO

GO was developed by the Gene Ontology Consortium to describe gene products using controlled and structured vocabulary and is divided into three categories: biological process, molecular function and cellular component.¹⁸ GO has been widely adopted by the life sciences community for gene function studies. Fisher’s exact test was applied to a gene set to calculate the significance

of a GO term, and the significant GO term sets were selected according to their *P*-values.⁷ The significant term sets of a disease gene set and a pathway make up their functional profiles, namely FPD and FPP, respectively.

Concordance scores between FPD and FPP

To quantify the functional association between a disease gene set and a pathway, we calculated the semantic similarity between an FPD and an FPP based on GO using information content (IC) theory.¹⁹ The IC of a GO term is defined as $IC(t) = -\log(p(t))$, where $p(t)$ is the number of genes annotated with the term t divided by the total number of genes annotated using GO. The semantic similarity for two GO terms can be measured by their shared IC, which is the most informative common ancestor (MICA) in the GO hierarchy. MICA is defined as $IC(t_{MICA}) = \max_{i(t_1, t_2)} [-\log p(t)]$. The measure was normalised to take into account the depth of the two GO terms.^{20,21} This normalised measure is given as follows:

$$\text{sim}(t_1, t_2) = \frac{2IC(t_{MICA})}{IC(t_1) + IC(t_2)} \quad (1)$$

For FPD and FPP, a similarity matrix $S = [s_{ij}]_{m \times n}$ was obtained, including the similarity scores between all terms in the FPD and FPP. We used the best-matched average measure (BMA) on matrix *S* to calculate the concordance score between an FPD and an FPP. BMA measurement finds the most similar GO term between the FPD and FPP for each GO term annotated to the FPP/FPD and then takes the average of these best matches.²² The concordance score is defined as follows:

$$\text{ConScore}(t_{\text{FPD}}, t_{\text{FPP}}) = \frac{1}{2} \left(\frac{1}{m} \sum_{i=1}^m \max_{1 \leq j \leq n} s_{ij} + \frac{1}{n} \sum_{j=1}^n \max_{1 \leq i \leq m} s_{ij} \right) \quad (2)$$

Recently, we demonstrated that it is practicable to calculate the functional similarity between two gene sets using semantic similarity measurement.²³ Thus, we could obtain the association between a disease gene set and a pathway by translating them into functional profiles and calculating a concordance score. After the pathways in a collection are prioritised according to their scores, plausible disease pathways are selected. A workflow diagram of this algorithm is illustrated in Figure 1. Note that a biological process is defined as a series of events accomplished by one or more ordered assemblies of molecular functions; thus, the ‘biological process’ category might relate most closely to our method for prioritising pathways. Hence, the ‘biological process’ category is used for gene annotation in this paper.

RESULTS

The performance of FPPP

To validate that the concordance score from the proposed method represents the biological truth, we have evaluated the ability of FPPP to identify known disease pathways. Known associations of 18 disease

gene sets from GAD with 26 disease pathways from KEGG are used as a benchmarking data set for our study.^{14,16} Each of the known disease–pathway association pairs is taken as one test case. To confirm that the concordance score is not randomly obtained, controls are artificially created to test the predictability of FPPP. We tested our method using random gene sets randomised both their pathway gene set randomisation and their disease gene set randomisation; we then computed empirical permutation *P*-values and plotted receiver-operating characteristic (ROC) curves.

First, for each of the 26 known pathways, 9999 pseudo-lists of disease genes were randomly selected from all disease genes in the GAD database, and the number of genes in each pseudo list was the same as the number in the true disease set. We added the true disease set for the pathway to the 9999 pseudo disease sets, resulting in a total of 10 000 sets. Next, we calculated the concordance scores between the 10 000 sets and the known pathway. We computed the empirical permutation *P*-value, a measure of how often a random pseudo set of disease genes yielded a concordance score that was equal or better than the observed concordance score for a disease. The empirical permutation *P*-value of 12 pathways was <0.0005 , and 19 pathways had *P*-values <0.005 (see Supplementary Table S2). We then calculated sensitivity and specificity for these pathways and plotted an ROC curve for 26 known disease–pathway pairs. The AUC value was as great as 95.02% (Figure 2a).

As another kind of test, we validated our method using the random pseudo pathways test. The genes of pseudo pathways were strictly limited to KEGG pathway genes, and the number of genes in each pathway was the same as in the true disease pathway. For each of the 18 diseases corresponding to 26 pathways, 10 000 pathways (consisting of one true disease pathway and 9999 pseudo pathways) were constructed for further testing. In spite of this stringent condition, 16 diseases yielded empirical permutation *P*-values <0.0005 , and 19 diseases had *P*-values <0.001 (see Supplementary Table S2). The AUC value of the rank of 18 diseases corresponding to 26 pathways was 96.73% (Figure 2b). The performance of our method is thus well validated by *P*-value and ROC curve analysis.

To further verify that the functional concordance score represents biological truth, we prioritised 186 pathways according to the concordance score for 18 diseases (see Supplementary Table S3). For the rank position of 18 known disease–pathway pairs, we also plotted an ROC curve, which yielded an AUC of 89.07% (Figure 2c). Next, we generated a detailed list of the top 20 pathways of 18 diseases. The top 20 pathways were considered disease-risk pathways. The 13 known disease pathways appeared among the set of risk pathways identified by our method. After prioritisation, all 18 diseases shared common-risk pathways, and they were associated with specific-risk pathways. There are 58 common pathways listed in the top 20 pathways of the 18 diseases. For example, the MAPK signalling pathway showed a strong association with all 18 diseases, and the cytokine–cytokine receptor interaction is related to 17 diseases, and apoptosis is related to 16 diseases. In contrast, the hypertrophic cardiomyopathy pathway only occurred in cardiomyopathy, and the type II diabetes mellitus pathway only occurred in diabetes. The scores between the diseases and pathways are displayed as a heat map (Figure 2d). These tests suggest that the proposed method has a powerful ability to predict true pathway prioritisations from among random controls, thereby inferring disease-risk pathways.

The robustness of FPPP

In many cases, a set of disease genes detected in experiments contain sources of noise, such as incomplete lists of disease genes and

false-positive genes. Therefore, robustness analysis of the methods for identifying disease-risk pathways based on a set of disease genes becomes important. Here, we studied the impact of noise on our method to test its robustness. In each of the known pathways, we introduced 10% random genes and repeated this analysis 100 times. The average value of the concordance score of known disease–pathway pairs with 10% noise was calculated. Similarly, we increased the proportion of noise in 10% increments and calculated the scores. We constructed curves of the concordance scores of 26 known disease–pathway pairs to assess the robustness of our method and the potential imprecision of the score. The concordance scores displayed low fluctuations even when the pathway contained up to 30% noise (see Supplementary Figure S2). These results indicate that our concordance score is relatively insensitive to noise and reliable for the prioritisation of disease pathways.

Prioritising the risk pathways of DCM

In this section, we show a case study of the FPPP method to test its reliability. The data set contains 403 genes differentially expressed in DCM derived from a gene expression profile using Affymetrix (Heidelberg, Germany) microarray chips.²⁴ DCM is a disease of the heart muscle that results in weakened contractions and poor pumping ability. Decreased heart function can affect the survival of patients. The details of this study can be found in the reference. Here, we analysed the pathway prioritisation of DCM by the FPPP method using 403 differentially expressed genes. On the basis of the concordance scores, the top 20 pathways are shown in Table 1. Three pathways related to heart disease were successfully identified, including arrhythmogenic right ventricular cardiomyopathy (ARVC), hypertrophic cardiomyopathy (HCM) and DCM. Using the FPPP method, regulation of the actin cytoskeleton, focal adhesion, adherens junctions, axon guidance and ECM receptor interactions were correlated with the pathogenesis of DCM. Previous research revealed that the actin cytoskeleton regulation pathway and the focal adhesion pathway have a pivotal role in cardiomyocyte survival and function. Our results found that DCM is consistent with dysfunction in the following pathways: actin cytoskeleton regulation, focal adhesion, gap junction, ECM receptor interaction and the adherens junction.^{25–31} A large number of signal transduction cascades have been implicated as critical regulators of cardiac hypertrophy, and the TGF beta signalling and MAPK signalling pathways were identified by our method.^{32–34} The other pathways in the top 20 were mostly disease pathways. Pathways such as axon guidance showed strong correlation with DCM based on their concordance scores by FPPP. Our method provides researchers with avenues for further study of the genetic basis of DCM. We believe that dysfunction of these pathways in combination lead to DCM. Further studies are needed to verify the relationships among these pathways.

The most widely used pathway analysis approach for identifying disease-related pathways is overrepresentation pathway analysis. To demonstrate the advantage of FPPP, we compared the results of FPPP with those of DAVID, which is widely used.¹² We input 403 differentially expressed genes of DCM and selected the Homo_sapiens data set as background. The statistically significant pathways were regarded as risk pathways if a false discovery rate of $q < 0.1$ using Fisher's exact test was achieved. We performed DAVID on the DCM genes and identified nine significant pathways. The risk pathways identified by both DAVID and FPPP were the regulation of the actin cytoskeleton, focal adhesion, ECM receptor interaction, the TGF beta signalling pathway and colorectal cancer. The known pathways of cardiomyopathy—ARVC, HCM and DCM—were all identified by our method, whereas none of them were identified by DAVID.

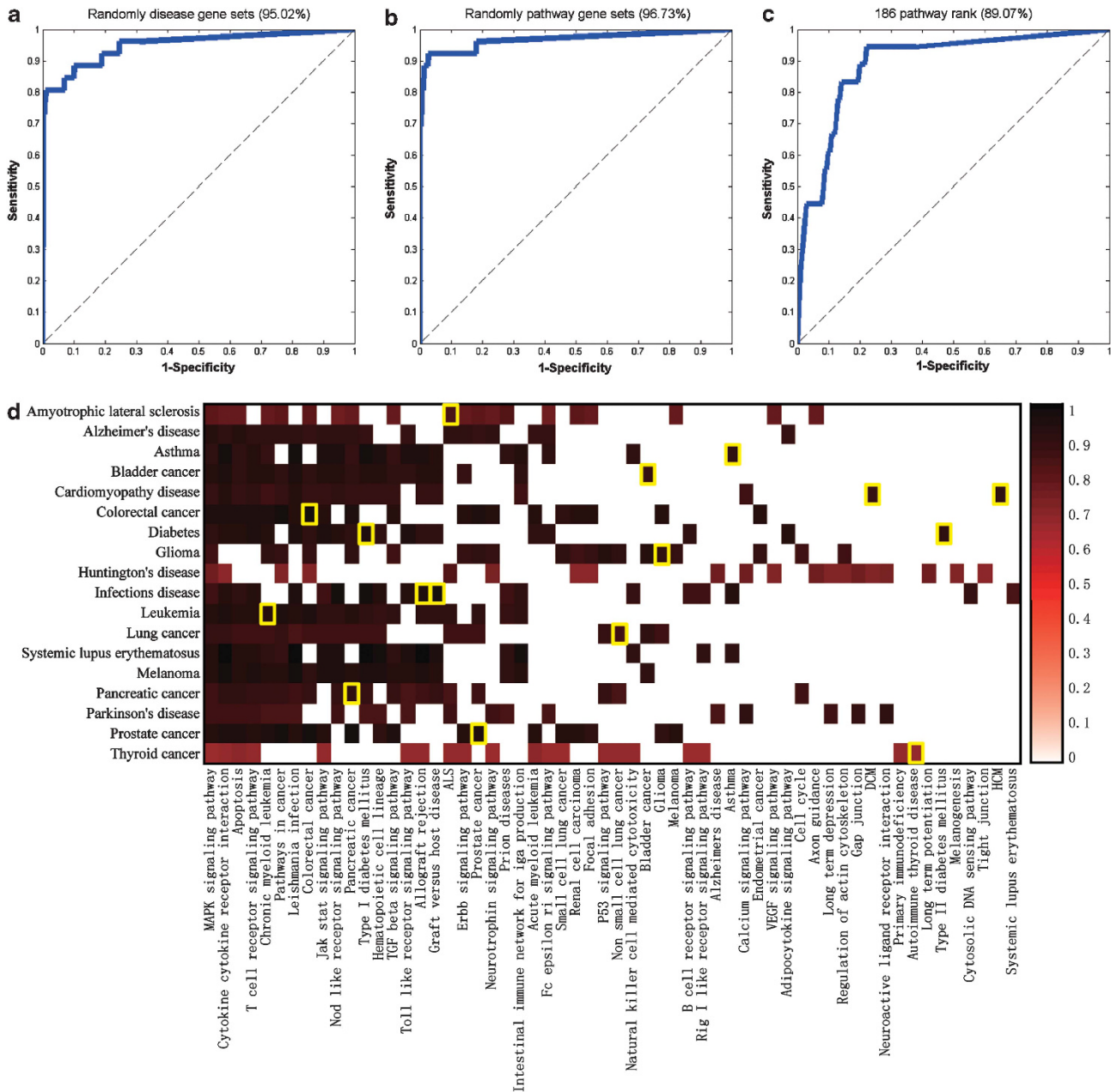


Figure 2 Validation results of FPPP. (a) ROC curves of known disease–pathway pairs in disease gene set randomisation. (b) ROC curves of known disease pathway pairs in pathway gene set randomisation. (c) ROC curves of known disease pathway pairs in 186 pathways. (d) The y axis represents the 18 diseases, and the x axis represents the 58 common pathways listed among the top 20 pathways of the 18 diseases. The concordance score is indicated by colour intensity, with black representing high values and white representing low values. The colour of each cell represents the concordance score of a disease (column) and a pathway (row), where black/white indicates a high/low concordance score. The regions outlined in yellow represent the scores of known disease–pathway associations.

In addition, another existing class of pathway analysis approaches mainly utilises information from the literature to identify biological pathways associated with gene expression changes.^{35–37} We compared our approach with an approach based on a gene expression data set. Here, we selected a widely used tool, the weighted global test tool.³⁷ We found that the known pathways of cardiomyopathy were not identified by the weighted global test.

At the same time, we made an elaborate comparison of three approaches using an acute myeloid leukaemia and acute lymphoblastic leukaemia data set.³⁸ The top 20 biological pathways for the FPPP

method and a global test approach, and 11 statistically significant pathways for DAVID, are listed in Supplementary Table S4. Compared with these existing approaches, the FPPP method provided better and more complementary insights for prioritising risk pathways. In addition, the prioritisations of two case data sets for the Reactome and Biocarta databases are listed in Supplementary Table S5.

The disease association landscape using FPPP

Research indicates that disease associations are caused by the dysfunction of several causal biological factors that are common to

multiple diseases. We assumed that the basis of associations between diseases potentially stems from the same or related biological pathways. Here, we used overlapping biological risk pathways among diseases to chart a genetic landscape of human disease. We first prioritised the pathways for 413 diseases from the GAD database¹⁶ using FPPP, thus yielding the disease-risk pathways for each disease (the top 20 pathways are considered as risk pathways). Next, we

Table 1 The top 20 of 186 pathways in DCM by the FPPP method and a conventional overrepresentation approach

Pathway name	Concordance	Enrichment analysis
	score	P-value
Regulation of actin cytoskeleton ^a	0.74048	0.016
ARVC ^b	0.71707	>0.1
Focal adhesion ^a	0.71219	1.00E-04
Adherens junction ^a	0.7098	>0.1
Axon guidance	0.70933	>0.1
ECM receptor interaction ^a	0.70617	9.70E-04
TGF beta signalling pathway ^a	0.70512	0.0053
Olfactory transduction	0.70234	>0.1
Pathways in cancer	0.70005	>0.1
Colorectal cancer	0.68693	0.055
HCM ^a	0.68507	>0.1
DCM ^a	0.68418	>0.1
Gap junction ^a	0.68067	>0.1
Pancreatic cancer	0.67671	>0.1
MAPK signalling pathway ^a	0.67478	>0.1
Fc gamma r-mediated phagocytosis	0.67467	>0.1
Prostate cancer	0.67031	>0.1
Melanoma	0.66898	>0.1
Renal cell carcinoma	0.66538	>0.1
Bladder cancer	0.66447	>0.1

Abbreviations: ARVC, arrhythmogenic right ventricular cardiomyopathy; DCM, dilated cardiomyopathy; HCM, hypertrophic cardiomyopathy; TGF, transforming growth factor.

^aDCM-related pathways documented and prioritised in the top 20 of 186 pathways.

^bKnown DCM pathways prioritised in the top 20 of 186 pathways.

computed the similarity scores between disease pairs using the number of pathway intersections and pathway unions, yielding a similarity score matrix. A two-way hierarchical clustering was performed on the matrix using Cluster+TreeView to reveal the modular organisation of human disease-disease relationships (Figure 3). Diseases clustered together are generally more similar and share more pathways. The disease landscape could systematically reveal the global modular organisation of disease association.

The modularity of the disease landscape is manifested as many isolated and highly scored modules, each comprising a set of diseases implicated in a set of genetically overlapping pathways. According to previous disease classifications, the disease modules are assigned to different markers, such as cancer, cardiovascular or immunity and infection. As shown in Figure 3, one disease class may be separated into several disease modules. For example, a module marked Cardiovascular1 is composed of myocardial infarction, restenosis, atherosclerosis, coronary-syndrome, blood, cardiovascular disease lacking more detailed labels, cardiovascular abnormality lacking more detailed labels, longevity, ischaemic-stroke, Alzheimer's disease and gestation-related conditions; most of these diseases or conditions are related to cardiovascular disease. Obviously, diseases in the module show high similarities and remarkably overlapping pathways, suggesting that they probably share underlying molecular mechanisms. The module marked Cardiovascular2 involves more cerebrovascular diseases, such as brain cancer, dementia, thrombophilia, cerebral amyloid angiopathy (senile plaques) and ischaemia. Interestingly, the Cardiovascular2 module shows only modest similarities, suggesting that diseases in the module have different pathogeneses. The particular diseases in each module are shown in Supplementary Table S6. We hope that the predicted disease landscape will facilitate future discovery of disease associations.

DISCUSSION

Investigation of the molecular basis of diseases is a major focus in genomics research. The occurrence of complex diseases results from the joint deregulation of many interactive genes rather than a mutation of an individual gene. Therefore, many researchers no

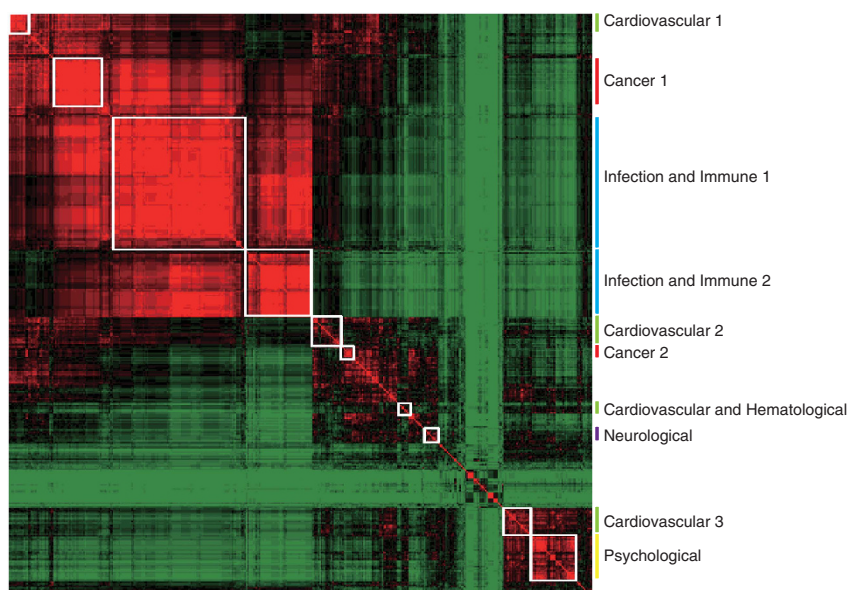


Figure 3 A cluster graph of disease associations with risk pathways from FPPP. The clustering result is displayed as a heat map, and the similarity score is indicated by colour intensity, with red representing high disease association and green representing low disease association.

longer study a single gene or protein but systematically and simultaneously analyse multiple genes or proteins. Compared with the analysis of an individual gene, pathway analysis might be a much more appropriate perspective from which to analyse disease biology because a pathway with internal relevance is more indicative of a certain function than a single gene. Pathway analysis may be much closer to the molecular basis of disease occurrence.

From this perspective, we propose an approach that ranks and infers disease-risk pathways based on a functional profiling method. Most existing pathway analysis approaches are based on gene overlap between pathways and diseases to identify disease-related biological pathways. We have recognised that non-overlapping genes might harbour indirect associations that could be useful for improving pathway analysis. Various factors may contribute to functional associations; these include substitutable and interactive genes. On the basis of the above notion, we are interested in studying the global function of all genes. The FPPP method is based on global functional profiling of all genes to identify disease-risk pathways. From a global and functional view, our approach may be closer to real biological pathways than approaches based on overlapping genes.

In addition, other types of existing pathway analysis approaches mainly utilise data from the literature to identify biological pathways associated with gene expression changes, taking into account all known genes for a given disease. Therefore, we compared our approach with these approaches using two expression data sets. Compared with existing approaches, the FPPP method could provide complementary insights to prioritise risk pathways. Evaluation of our method illustrates the power of inferring disease-risk pathways by examining known disease–pathway associations. It also showed a great ability to resist noise when ranking disease pathways. Finally, the identified risk pathways of 413 diseases were used to depict a disease similarity landscape that reveals the global modular organisation of disease associations.

Notably, FPPP is based on functional profiles that depend on GO annotation. Although the GO database is the most widely used and provides a large amount of gene annotation information, it remains an imperfect system and needs further improvement. Thus, the incompleteness of the GO database can influence our results. Improvements in GO annotations will allow our approach to produce better results for pathway prioritisation. In addition, many other biological resources, such as larger sets of expression profiles, are available. These biological resources may be used to help identify complex functional associations and can further contribute to advances in the study of complex molecular disease mechanisms.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (grant nos. 61073136, 91129710, 61170154 and 31100948), the Specialized Research Fund for the Doctoral Program of Higher Education of China (grant nos. 20102307110022) and the Science Foundation of Heilongjiang Province (grant nos. JC200711, ZD200816-01 and D201114).

- 3 Taylor BS, Schultz N, Hieronymus H *et al*: Integrative genomic profiling of human prostate cancer. *Cancer Cell* 2010; **18**: 11–22.
- 4 Iliopoulos D, Jaeger SA, Hirsch HA, Bulyk ML, Struhl K: STAT3 activation of miR-21 and miR-181b-1 via PTEN and CYLD are part of the epigenetic switch linking inflammation to cancer. *Mol Cell* 2010; **39**: 493–506.
- 5 Efferson CL, Winkelmann CT, Ware C *et al*: Downregulation of Notch pathway by a gamma-secretase inhibitor attenuates AKT/mammalian target of rapamycin signaling and glucose uptake in an ERBB2 transgenic breast cancer model. *Cancer Res* 2010; **70**: 2476–2484.
- 6 Goeman JJ, Buhlmann P: Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 2007; **23**: 980–987.
- 7 Huang da W, Sherman BT, Lempicki RA: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009; **37**: 1–13.
- 8 Zhou X, Su Z: EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agricultural species. *BMC Genomics* 2007; **8**: 246.
- 9 Zheng Q, Wang XJ: GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res* 2008; **36**: W358–W363.
- 10 Mlecnik B, Scheidele M, Hackl H, Hartler J, Sanchez-Cabo F, Trajanoski Z: PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res* 2005; **33**: W633–W637.
- 11 Dennis Jr. G, Sherman BT, Hosack DA *et al*: DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol* 2003; **4**: P3.
- 12 Huang da W, Sherman BT, Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009; **4**: 44–57.
- 13 Subramanian A, Tamayo P, Mootha VK *et al*: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005; **102**: 15545–15550.
- 14 Kanehisa M: The KEGG database. *Novartis Found Symp* 2002; **247**: 91–101, discussion 101–103, 119–128, 244–152.
- 15 Vastrik I, D'Eustachio P, Schmidt E *et al*: Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* 2007; **8**: R39.
- 16 Becker KG, Barnes KC, Bright TJ, Wang SA: The genetic association database. *Nat Genet* 2004; **36**: 431–432.
- 17 Li X, Li C, Shang D *et al*: The implications of relationships between human diseases and metabolic subpathways. *PLoS One* 2011; **6**: e21131.
- 18 Ashburner M, Ball CA, Blake JA *et al*: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; **25**: 25–29.
- 19 Pesquita C, Faria D, Falcao AO, Lord P, Couto FM: Semantic similarity in biomedical ontologies. *PLoS Comput Biol* 2009; **5**: e1000443.
- 20 Resnik P: Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J Artif Intell Res*, **11**: 1999; 95–130.
- 21 Lin D: An information-theoretic definition of similarity. *In Proc of the 15th Int Conf on Machine Learning* 1998; 296–304.
- 22 Schlicker A, Domingues FS, Rahnenfuhrer J, Lengauer T: A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* 2006; **7**: 302.
- 23 Lv S, Li Y, Wang Q *et al*: A novel method to quantify gene set functional association based on gene ontology. *J R Soc Interface* 2011; **9**: 1063–1072.
- 24 Barth AS, Kuner R, Bunes A *et al*: Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies. *J Am Coll Cardiol* 2006; **48**: 1610–1617.
- 25 Melendez J, Welch S, Schaefer E *et al*: Activation of pyk2/related focal adhesion tyrosine kinase and focal adhesion kinase in cardiac remodeling. *J Biol Chem* 2002; **277**: 45203–45210.
- 26 Zhang H, Zhang A, Guo C *et al*: S-diclofenac protects against doxorubicin-induced cardiomyopathy in mice via ameliorating cardiac gap junction remodeling. *PLoS One* 2011; **6**: e26441.
- 27 Sheikh F, Chen Y, Liang X *et al*: alpha-E-catenin inactivation disrupts the cardiomyocyte adherens junction, resulting in cardiomyopathy and susceptibility to wall rupture. *Circulation* 2006; **114**: 1046–1055.
- 28 Connelly KA, Kelly DJ, Zhang Y *et al*: Inhibition of protein kinase C-beta by ruboxistaurin preserves cardiac function and reduces extracellular matrix production in diabetic cardiomyopathy. *Circ Heart Fail* 2009; **2**: 129–137.
- 29 Eigenthaler M, Engelhardt S, Schinckel B *et al*: Disruption of cardiac Ena-VASP protein localization in intercalated disks causes dilated cardiomyopathy. *Am J Physiol Heart Circ Physiol* 2003; **285**: H2471–H2481.
- 30 Hammer E, Goritzka M, Ameling S *et al*: Characterization of the human myocardial proteome in inflammatory dilated cardiomyopathy by label-free quantitative shotgun proteomics of heart biopsies. *J Proteome Res* 2011; **10**: 2161–2171.
- 31 Sun YM, Zhang Y, Zeng LQ *et al*: Broad profiling of DNA-binding transcription factor activities improves regulatory network construction in adult mouse tissues. *J Proteome Res* 2008; **7**: 4455–4464.
- 32 Molkenin JD: Calcineurin-NFAT signaling regulates the cardiac hypertrophic response in coordination with the MAPKs. *Cardiovasc Res* 2004; **63**: 467–475.
- 33 Miao Y, Zhang W, Zhong M, Ma X, Qi TG, Sun H: Activation of transforming growth factor-beta1/Smads signal pathway in diabetic cardiomyopathy and effects of valsartan thereon: experiment with rats. *Zhonghua Yi Xue Za Zhi* 2007; **87**: 366–370.

- 1 Kelley R, Ideker T: Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* 2005; **23**: 561–566.
- 2 Kim T, Yoon J, Cho H *et al*: Downregulation of lipopolysaccharide response in *Drosophila* by negative crosstalk between the AP1 and NF-kappaB signaling modules. *Nat Immunol* 2005; **6**: 211–218.

- 34 Heineke J, Molkenin JD: Regulation of cardiac hypertrophy by intracellular signalling pathways. *Nat Rev Mol Cell Biol* 2006; **7**: 589–600.
- 35 Leong HS, Kipling D: Text-based over-representation analysis of microarray gene lists with annotation bias. *Nucleic Acids Res* 2009; **37**: e79.
- 36 Minguez P, Al-Shahrour F, Montaner D, Dopazo J: Functional profiling of microarray experiments using text-mining derived bioentities. *Bioinformatics* 2007; **23**: 3098–3099.
- 37 Jelier R, Goeman JJ, Hettne KM, Schuemie MJ, den Dunnen JT, t Hoen PA: Literature-aided interpretation of gene expression data with the weighted global test. *Brief Bioinform* 2011; **12**: 518–529.
- 38 Golub TR, Slonim DK, Tamayo P *et al*: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; **286**: 531–537.

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