BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 2 2009, pages 237–242
doi:10.1093/bioinformatics/btn613

Genetics and population analysis

Prioritizing risk pathways: a novel association approach to searching for disease pathways fusing SNPs and pathways

Lina Chen*,†, Liangcai Zhang†, Yan Zhao†, Liangde Xu†, Yukui Shang, Qian Wang, Wan Li, Hong Wang and Xia Li*,†

College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

Received on August 1, 2008; revised on October 27, 2008; accepted on November 21, 2008

Advance Access publication November 24, 2008

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Complex diseases are generally thought to be under the influence of one or more mutated risk genes as well as genetic and environmental factors. Many traditional methods have been developed to identify susceptibility genes assuming a single-gene disease model ('single-locus methods'). Pathway-based approaches, combined with traditional methods, consider the joint effects of genetic factor and biologic network context. With the accumulation of high-throughput SNP datasets and human biologic pathways, it becomes feasible to search for risk pathways associated with complex diseases using bioinformatics methods. By analyzing the contribution of genetic factor and biologic network context in KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways, we proposed an approach to prioritize risk pathways for complex diseases: Prioritizing Risk Pathways fusing SNPs and pathways (PRP). A risk-scoring (RS) measurement was used to prioritize risk biologic pathways. This could help to demonstrate the pathogenesis of complex diseases from a new perspective and provide new hypotheses. We introduced this approach to five complex diseases and found that these five diseases not only share common risk pathways, but also have their specific risk pathways, which is verified by literature retrieval.

Availability: Genotype frequencies of five case–control samples were downloaded from the WTCCC online system and the address is https://www.wtccc.org.uk/info/access_to_data_samples.shtml

Contact: chenlina@ems.hrbmu.edu.cn; lixia@hrbmu.edu.cn

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

1 INTRODUCTION

Complex diseases are known to arise from one or more mutated genes as well as genetic and environmental factors. One major challenge of the post-genomic era is to find the genes at risk, identify their functions and develop new techniques for testing, diagnosis and treatment (Mocellin *et al*., 2004). The complexity of these diseases cannot be interpreted by a single gene product or the behavior of a single specific pathway. Their pathogenesis should be interpreted with the effects of genetic and environmental factors together (Kelley and Ideker, 2005). When case–control datasets of complex diseases are available, genome-wide association studies (GWA) have greater power to detect genetic variants, even if many markers are tested across the genome. All published GWA studies have led to the discovery of novel genes for several complex diseases, but there are limitations (Wang *et al*., 2007). First, genetic variants that confer a small risk of disease and contain potential biologic importance are likely to be missed in the 'most significant SNPs/genes' approach after adjustment for multiple testing. Second, even those variants that confer a larger effect may not always be tested, especially if the sample size is small. Discovering new strategies of how to avoid the limitations of studies of complex diseases is vital. Considering the effects of biologic network context systematically is also important. With ongoing development of databases for metabolic pathways, scientists could effectively research complex diseases against a background of the biologic metabolic environment. The KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway is a network of molecular wiring diagrams of interaction and reaction, and it could reflect the complexity and diversity of the metabolic environment (Ogata *et al*., 2000). With the accumulation of high-throughput datasets, studying the pathogenesis of complex diseases from the viewpoints of genetic factor and KEGG network context systematically is feasible.

Focusing on genetic factors and KEGG network context, we calculated an approach to find risk pathways associated with complex diseases: prioritizing risk pathways fusing SNPs and pathways (PRP). By appropriately quantifying genetic factor and biologic network context and integrating them, we worked out a risk-scoring (*RS*) measurement to find intimate biologic pathways that could clearly demonstrate the pathogenesis of complex diseases from a new viewpoint.

This method was introduced for five complex diseases: bipolar disorder (BD), coronary artery disease (CAD), hypertension (HT), type-2 diabetes (T2D) and inflammatory bowel disease (CD). Compared with other enrichment methods, the PRP method could return richer information. This method will be introduced to other complex diseases to provide additional insights into their pathogenesis.

2 DATA SOURCE

Genotype frequencies of tested SNPs of case–control samples were downloaded from the Wellcome Trust Case Control Consortium

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, first four authors and the last author should be regarded as joint First Authors.

(WTCCC) online system from the following samples using the 500K Affymetrix chip: 2000 samples each from the five disease collections, i.e. BD, CAD, HT, T2D and CD (WTCCC, 2007). To complete the PRP method, we preprocessed the genotype frequencies to allele frequencies for each SNP. About 200 human pathways were obtained from the KEGG database (Kanehisa, 2002). Location information of the human genes was acquired from the NCBI genome database.

3 METHODS

3.1 SNP significance analysis and risk evaluation

Data (D) depicted the frequency of cases and controls for each of the alleles at a SNP locus (Table 1).

According to matrix D, the statistics χ^2 were computed first, then the corresponding probability P calculated, and each SNP was assigned a specific P-value. Under the significance level of 0.05, a SNP set was preliminarily screened for the following study. The *Pearson*-χ^2 formula was as follows:

$$\chi^2 = \frac{(ad-bc)^2\,(a+b+c+d)}{(a+b)\,(b+d)\,(d+c)\,(c+a)}$$

A *risk* statistic can be acquired for each SNP if its significance level meets the threshold of 0.05 (Tian *et al*., 2008), and the formula is:

$$risk = \begin{cases} \log \frac{f_{case(A)}}{f_{control(A)}} & f_{case(A)} > f_{control(A)} > 0 \\ \log \frac{f_{case(B)}}{f_{control(B)}} & f_{case(B)} > f_{control(B)} > 0 \\ 0 \quad \text{others} & \end{cases}$$

Where $f_{control}(A)$, $f_{control}(B)$, $f_{case}(A)$ *and* $f_{case}(B)$ are the frequencies of controls and cases at the allele A and B, respectively.

If the P-value of the selected SNP was not significant, we set its *risk* value to be 0 directly. For each SNP, the *risk* value was introduced to depict its relationship with complex disease. The *risk* value could, in some sense, reflect the genetic effects of complex disease.

3.2 Pathway reconstruction and biologic network context analysis

The KEGG pathway is a network in which a node represents the metabolite and one edge represents an enzyme or a gene cluster. First, each KEGG pathway was translated into a graph whose edge was a metabolite and the node was an enzyme or a gene cluster. An 'entity-relationship graph' K of enzymes/genes was produced. A sample pathway is shown in Figure 1 (Ogata *et al*., 2000).

The graph K describes the relationship between gene clusters well. The degree attribution of the node, the number of its adjacent edges reflect the diversity of the metabolites involved in different biologic reactions in the primary pathway (i.e. the degree could concisely demonstrate the complexity of the metabolic network context). Here, for each node, the biologic network context refers to its degree attribution in the reconstructed pathway.

Table 1. Frequency of cases and controls for each of two alleles at a SNP locus

Allele	A	B
Case	a	b
Control	c	d

In the table, a-d represents the frequency of the specific allele A or B, respectively.

3.3 SNP screening and evaluation of gene risk

For each SNP V_i $(i = 1,\ldots,L$, where L is the total number of SNPs in the GWA study), we calculated its *risk* statistic (seen in Section 3.1). We then associated gene g_t $(t = 1,\ldots,T$, where T is the number of all the pathway genes) with SNP V_i if this SNP is located within g_t or if g_t is the closest to V_i. SNPs that are 500 kb away from any gene were considered because most enhancers and repressors are <500 kb away from genes, and most linkage disequilibrium blocks are <500 kb (Wang *et al*., 2007).

For each gene g_t, we assigned the highest *risk* of all the SNPs located within g_t as its genetic statistic. The corresponding formula was:

$$Risk(g_t) = \max_{1 \leqslant i \leqslant L_t} \{risk_1, risk_2, \ldots, risk_i, \ldots, risk_{L_t}\}$$

where $risk_i$ is the *risk* value of the i-th SNP located in g_t, and L_t is the total number of SNPs mapped to g_t.

3.4 Integration of genetic factor and biologic network context

In general, the phenotypes of complex diseases are thought to be under the joint influence of genetic and environmental factors. Quantifying genetic factor and biologic network context and integrating these two aspects has enabled an integrated measurement to be calculated to clearly reveal the pathogenesis of complex disease. The relationship between phenotype and corresponding influential factors is shown in Figure 2.

Under the specific background of pathway $k_i (i = 1,\ldots,U$, where U is the number of human pathways in KEGG database), for each gene cluster s_j $(j = 1,\ldots,N_i$, where N_i is the count of gene clusters in pathway $k_i)$, the biologic network context $E(s_j, k_i)$ of the gene cluster s_j was measured according to its degree attribution in the reconstructed background pathway k_i. $E(s_j, k_i)$, the degree of gene cluster s_j, was the number of its adjacent edges in the reconstructed pathway k_i (see Section 3.2). The genetic factor $G(P, s_j)$ of the gene cluster s_j was quantified with an appropriate average measurement for

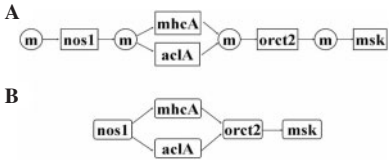

Fig. 1. Sample pathway. (**A**) Sample pathway of KEGG in which rectangles represent reactions (labeled with names of genes encoding for their catalyzing enzymes) and circles labeled with the letter 'm' represent metabolites (which are not explicitly named here for simplicity) being consumed and/or produced. (**B**) Depicts a sample-reconstructed graph K of the original pathway.

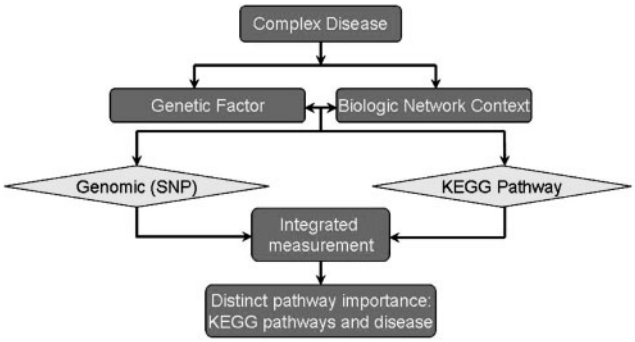

Fig. 2. Integration of genetic factor and biologic network context.

phenotype P (M_j is the size of s_j), and its formula was:

$$G(P,s_j)=\frac{1}{M_j}\sum_{t=1}^{M_j}Risk(g_t)$$

Using s_j as the bridge, the RS value $preRS(P,k_i)$ between phenotype P and pathway k_i was:

$$preRS(P,k_i)=\sum_{j=1}^{N_i}\{G(P,s_j)\times E(s_j,k_i)\} \quad (1)$$

To identify relative RSs between pathways, we promoted the standardization to make them zero dimension and comparable. Obviously, the minimum of RS will be always 0 for each phenotype (similar to real observations). We evaluated the final RS of k_i, $RS(P, k_i)$, with relative value by dividing the deviation between $preRS(P, k_i)$ and the minimum with the deviation between the maximum and minimum value. And finally the standardized value $RS(P, k_i)$ was calculated as:

$$RS(P,k_i)=\frac{preRS(P,k_i)}{\max\limits_{1\leqslant i\leqslant U}\{preRS(P,k_i)\}} \quad (2)$$

For example:

$$RS(P,k_i)=\frac{\sum\limits_{j=1}^{N_i}\{G(P,s_j)\times E(s_j,k_i)\}}{\max\limits_{1\leqslant i\leqslant U}\sum\limits_{j=1}^{N_i}\{G(P,s_j)\times E(s_j,k_i)\}} \quad (3)$$

3.5 Algorithm of PRP method

Comprehensive consideration of genetic factor and biologic network context involved a new measurement of RS to depict the association between biologic pathways and the eventual-specific phenotype (Fig. 3).

The steps of the PRP method are shown below.

(i) Calculate *risk* values for all the SNPs whose *P*-values meet the threshold of 0.05 according to the *Pearson*-χ^2 statistic.

(ii) Reconstruct KEGG pathways and compute the degree attribution, $E(s_j,k_i)$, for the gene cluster s_j in pathway k_i.

(iii) Screen and map SNPs to the corresponding genes in the reconstructed network k_i.

(a) Map all the SNPs in step (i) to the corresponding g_t that are located <500 kb away from g_t.

(b) Select the maximum *risk* value $Risk(g_t)$ as the genetic statistic value for g_t.

(iv) Compute $RS(P,k_i)$ for the pathway k_i.

(v) For $i=1$ to U, repeat step (iv), compute and prioritize all the *RS* values.

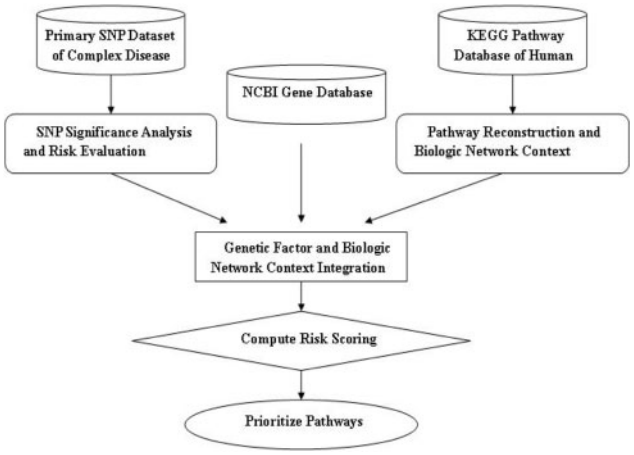

Fig. 3. Algorithm of the PRP method.

3.6 Randomization tests

Owing to the complex nature of the data—potentially genetic locus with differential significance in the pathway context—genotype frequencies randomization tests were introduced to estimate the stability of the PRP prioritization method. In this case, all the corresponding genotype frequencies were shuffled. This approach retained the structure of the reconstructed pathways and randomized only the frequency of all the genotypes.

Another randomization test should be carried out for each dataset, i.e. randomization of the degree attribution of each gene set in the reconstructed pathway. This leads to similar results to the first randomization (data not shown).

4 RESULTS

4.1 Risk pathways prioritization using PRP method

To study the association between pathways and diseases, the PRP method comprehensively considers the effects of genetic factor and biologic network context on complex diseases. Using the PRP method, RS of each pathway was calculated for each disease. For all five diseases, scoring information of the top 10 biologic pathways is shown in Figure 4.

According to the PRP method, all five chronic inflammatory diseases shared common risk pathways, and they had specific risk pathways. Purine metabolism, arachidonic acid metabolism and the MAPK signaling pathway showed a strong association with all five diseases (see Table 2 and Supplementary Table 1). There were high *RS* scores in the phosphatidylinositol signaling system and focal adhesion pathway for BD, CAD and CD (see Supplementary Table 2). For BD, CAD and T2D, the *RS* scores of WNT and insulin signaling pathways ranked high simultaneously.

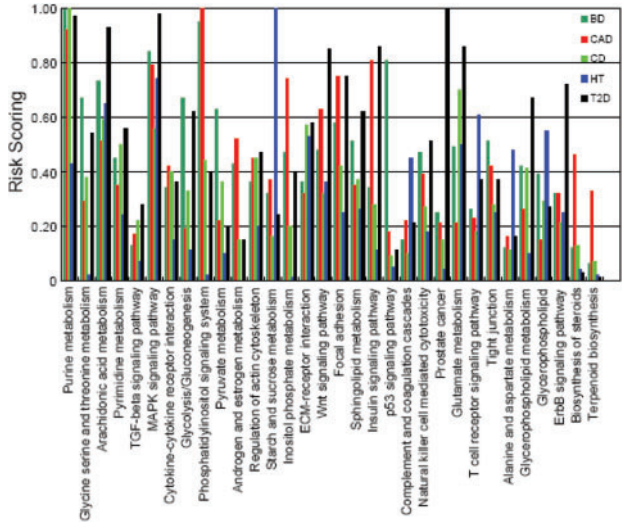

Fig. 4. *RS* scoring chart of the top 10 pathways associated with five complex diseases.

Table 2. *RS* scores and modular analysis for the top 10 risk pathways of BD

Pathway name	*RS*	Module	No.[a]	Robustness[b]
Purine metabolism[c]	1.00	M	233	Y
Phosphatidylinositol signaling system	0.95	EIP	27	Y
MAPK signaling pathway[c]	0.84	EIP	6	Y
p53 signaling pathway	0.81	CP	3	Y
Arachidonic acid metabolism[c]	0.73	M	40	Y
Glycolysis/Gluconeogenesis	0.67	M	3	Y
Glycine, serine and threonine metabolism	0.67	M	120	Y
Pyruvate metabolism	0.63	M	0	Y
Focal adhesion	0.58	CP	39	Y
Tight junction	0.51	CP	0	Y

[a]Numbers that verify the relationship between pathways and corresponding diseases in the literature. The RS column is the risk score of the PRP method. [b]Y/N refers the pathway is robust or not after one hundred randomization. The RS scores and modular analysis for the top 10 risk pathways are illustrated in the Supplementary Table 1 for the other four diseases (CAD, HT, CD and T2D). [c]Mutual risk pathways for five diseases.

In addition to T2D, HT and CD, all the risk scores appeared to be high in the pathway of glutamate metabolism. HT and CD got higher scores in the ECM–receptor interaction pathway. Each of the five complex diseases had specific risk pathways. The specific risk pathways associated with BD consist of glycine, serine and threonine metabolism, glycolysis/gluconeogenesis, pyruvate metabolism, p53 signaling pathway and tight junction. The specific pathways of CAD contained androgen and estrogen metabolism, biosynthesis of steroids and inositol phosphate metabolism. CD showed a high association with pyrimidine metabolism, regulation of the actin cytoskeleton and alanine and aspartate metabolism. The specific pathways for HT were starch and sucrose metabolism, complement and coagulation cascades, the T-cell receptor signaling pathway and glycerolipid metabolism. Interestingly, our approach found that T2D showed a higher association with prostate cancer and the ErbB signaling pathway.

4.2 Randomization results and modular clustering of high-risk pathways

We undertook randomization tests for our source datasets to examine the robustness of the PRP method. Unlike common permutation, permutation of the frequency of each SNP label and recalculation of its statistic values for hundreds or thousands of SNPs in the GWA analysis is a very computationally expensive process. One-hundred randomization tests were carried out for BD (Fig. 5). And the similar randomization results for the other four disease datasets could be found in Supplementary Figure 1.

Risk scores from the PRP method were generally higher than those of randomization tests (Fig. 5). For each pathway, we defined it as a risk pathway if its *RS* value of the PRP method was always >95% quantile of its random values. In the following analysis, we only consider pathways if their true risk scores were higher than their 95% quantile of all random scores.

Human pathways are classified into six categories in the KEGG pathway database: metabolism (M), cellular process (CP), environmental information process (EIP), genomic information

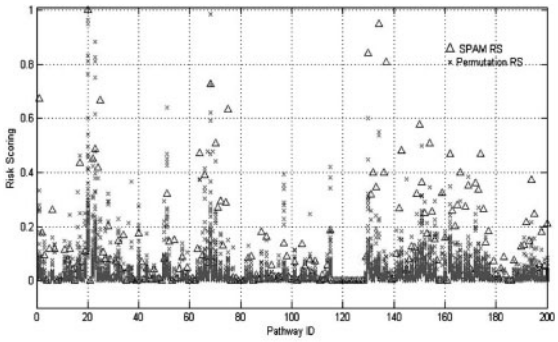

Fig. 5. Randomization tests results for BD. The *x*-axis represents the ordinal number of KEGG pathways; the *y*-axis is the *RS* scores of the PRP method (labeled by triangle in blue) and 100 times randomization tests (labeled by cross mark in red).

(GI), human disease (HD) and drugs development (DD) (detailed information of these categories could be acquired at KEGG online database and the corresponding website is http://www.genome.jp/kegg/pathway.html). Using these categories, all pathways were classified into the specific categories for the five complex diseases. All the high-risk pathways were concentrated in M, EIP and CP modules (Table 2).

4.3 Verification

To confirm the results of the PRP method, we used the PUBMED retrieval module of NCBI to verify the relationship between five complex diseases and the corresponding risk pathways by literature retrieval: our results were supported by many related studies.

BD (manic depressive illness) refers to an episodic recurrent pathological disturbance in mood ranging from extreme elation or mania to severe depression and usually accompanied by disturbances in thinking and behavior: psychotic features (delusions and hallucinations) often occur. Pathogenesis is poorly understood but there are robust evidences in substantial literatures. Using the PRP method, it was found that purine metabolism, phosphatidylinositol signaling system, MAPK signaling pathway, p53 signaling pathway, arachidonic acid metabolism, WNT signaling pathway and so on were correlated with the pathogenesis of BD (Table 2). Oral choline supplementation resulted in a significant decrease in brain purine levels through purine metabolism in lithium-treated patients with BD. This result is consistent with mitochondrial dysfunction in BD inadequately meeting the demand for increased ATP production as exogenous oral choline administration increases membrane phospholipid synthesis of phosphatidylinositol signaling system (Kim *et al*., 2005; Lyoo *et al*., 2003; Montezinho *et al*., 2007; Sjoholt *et al*., 2004). The activity of natural killer (NK) cells mediated by NK cell activity pathway turned out to be a significant difference in case–control experiment of patients with BD (Abeer *et al*., 2006). The cadherin gene FAT and its protein partners in the WNT pathway may be components of a molecular pathway involved in susceptibility to BD (Blair *et al*., 2006). The therapeutic effect of valproate on BD involves interfering with the metabolism of arachidonate and inositol through signaling systems such as the WNT signaling pathway (Bezchlibnyk and Young, 2002; Gould and Manji, 2002; Rosenberg, 2007).

For the other diseases (CAD, CD, HT and T2D), corresponding evidences could be found to verify the relationship between higher risk pathways and each phenotype in Supplementary Table 3. Using the PRP method, we found that androgen and estrogen metabolism is also a high-risk pathway associated with CAD. Mohamad *et al*. (2006) illustrated that in young women, in the presence of coronary risk factors and a normal level of serum estrogen, the high levels of serum free testosterone and low levels of serum sex hormone-binding globulin are associated with the development of atherosclerosis and increased incidence of CAD. Imig (2006) reported that arachidonic acid metabolites are vital for the appropriate control of renal haemodynamics and, if not properly controlled, can contribute to renal vascular injury and end-stage renal disease related to HT and T2D; renal vascular eicosanoids have important roles in the pathogenesis of renal disease related to HT, T2D, metabolic syndrome and acute renal failure. An interesting result from PRP analysis demonstrated that T2D is strongly associated with prostate cancer, and similar result was found that patients with T2D are at a decreased risk of prostate cancer (Fukui *et al*., 2008). This might indicate combined studies between phenotypes, which might contribute to the therapy of complex diseases. Further studies might be needed to verify the inner relationship.

In addition, we found that all five diseases shared common risk pathways from the results of PRP method. Literatures about the separate effect of these three metabolic processes on each phenotype could be found in Supplementary Table 3, while the underlying importance of mutual occurrence for all these risk pathways are still not known.

4.4 Comparison PRP method with other approaches

4.4.1 Enrichment analysis based on hypergeometric testing We carried out GWA studies to find the gene sets for all five diseases. The threshold of significant *P*-value and multiple testing (FDR) were set at 0.05. Based on genetic factors only, risk genes, which were selected by GWA studies and traditional multiple testing, were annotated in pathways (Supplementary Table 2). A traditional enrichment analysis based on hypergeometric distribution was carried out in KEGG pathways for all the risk genes, and the corresponding formula was:

$$p(k_i) = 1 - \frac{C_M^m C_{N-M}^{n-m}}{C_N^n}$$

where, N is the number of all the background pathway genes, n is the number of risk genes out of FDR multiple testing, M is the total number of genes in pathway k_i, while m is the overlap number of risk genes and pathway k_i genes.

For each disease, the distribution information of risk genes is illustrated in Supplementary Table 2. After hypergeometric testing, enrichment *P*-values of risk genes are not significant and there are no significant pathways as we expected. In one way, we could conclude that if pathways are specified, traditional enrichment analysis based on hyper geometric distribution may not return the results of the association between complex diseases and the pathway sets because only a handful of genes are significant in each pathway and no enough genes to do the statistics for each disease.

Table 3. Risk genes distribution results of traditional enrichment methods for BD

Pathway name	GIs[a]	Hypergeometric testing *P*-value	GSEA *P*-value
Purine metabolism	7	>0.05	0.000723
Phosphatidylinositol signaling system	3	>0.05	>0.05
MAPK signaling pathway	18	>0.05	0.000473
p53 signaling pathway	0	>0.05	>0.05
Arachidonic acid metabolism	6	>0.05	>0.05
Glycolysis/Gluconeogenesis	5	>0.05	>0.05
Glycine, serine and threonine metabolism	0	>0.05	>0.05
Pyruvate metabolism	3	>0.05	0.000473
Focal adhesion	8	>0.05	>0.05
Tight junction	7	>0.05	0.006743

[a]Number of significant genes in GWA studies (FDR < 0.05). The risk gene distribution results for CAD, HT, CD and T2D are illustrated in Supplementaary Table 2.

4.4.2 Gene set enrichment analysis Gene set enrichment analysis (GSEA) is a method based on functional categories. It considers the genetic factor associated with complex diseases and the statistically significant accumulations of genes that belong to a functional category (Backes *et al*., 2007). We used the GeneTrail online application to carry out the GSEA (FDR = 0.05) of risk genes on KEGG pathways (Table 3 and Supplementary Table 2). Compared with GSEA, the PRP method could return identical results, from which more important information can be garnered.

4.4.3 Comparison between the PRP method and traditional approaches According to the gene distribution in Table 3 and Supplementary Table 2, all the risk genes of the five diseases are predominantly in the EIP and CP modules, and the number of genes in corresponding pathways appear to be higher, which is consistent with results of the PRP method. Using hypergeometric testing method, there may be no hints for association between BD, pyrimidine metabolism and glycine, serine and threonine metabolism; and there is no clue for the relationship between T2D and glutamate metabolism. According to these two methods above, it could not show that arachidonic acid metabolism and the biosynthesis of steroids are related to CAD. And there is no hint for the association between T2D and prostate cancer, and no clue for the risk relationship between arachidonic acid metabolism and complex diseases such as HT and CAD (Supplementary Table 2).

The PRP method is different from traditional studies. The PRP method considered risk genes and their underlying biologic context, and dedicates to identify risk pathways of complex diseases. With risk genes and their biologic network context considered, the PRP method could not only find out the pathways that the most significant genes gather in, but also sensitively return the specific pathways, where fewer most significant genes locate. However, traditional methods dedicate to identify the most significant SNPs/genes for each disease in their SNP significant analyses. Complex diseases are thought to be caused by multiple risk genes mutually rather than most significant ones only. Compared with traditional methods, the PRP method could provide additional insights into the pathogenesis of complex diseases in the form of pathway or network.

5 DISCUSSION

Complex diseases are thought to be caused by multiple genetic and environmental factors. Their pathogenesis may be accompanied with the dysfunction of several metabolic pathways. The PRP method considers the genetic factor and biologic network context and returns more affluent information under the same biologic background. Based on genetic factors only, traditional approaches ignore some important biologic factors. As to the risk genes from traditional GWA and multiple testing, they are predominantly distributed not only in environmental information process and cellular process for all the five complex diseases (BD, T2D, CAD, CD and HT), but also in the metabolism category. Through literature annotation, the risk pathways associated with diseases are well verified, but studies about environmental information processes appear to be fewer. This suggests that the pathogenesis of complex diseases could be revealed from different perspectives and multiple factors.

Having obtained comprehensive datasets, we will carry out in-depth studies on the mutual effects of multiple factors, and how to use the PRP algorithm to reveal the pathogenesis of complex diseases. When researching complex diseases, investigators should not only focus on the effect of genetic factors, but also pay more attention to environmental factors. This may be an urgent problem to tackle for systems biology scientists.

It was proved that the PRP method could select risk pathways associated with complex diseases. This method was found to be feasible, and afforded more clues than traditional methods.

The PRP method would make researchers of complex diseases extend their consideration from genetic factors only to the combination of genetic factor and biologic network context to explain the mechanism in the pathogenesis. The PRP method could show good results in determining biologic pathways associated with complex diseases. In this study, we considered only the specific metabolism environmental impacts. If transcriptional and proteome datasets are accumulated and machine learning methods introduced, the PRP method could be used to reflect joint effects on the pathogenesis of complex diseases better from multiple perspectives.

ACKNOWLEDGEMENTS

We thank the Wellcome Trust Case Control Consortium (WTCCC) and his laboratory for the generosity of providing us with the genotype frequence of SNP data.

Funding: National High Tech Development Project of China, the 863 Program (Grant No. 2007AA02Z329); the National Natural Science Foundation of China (Grant Nos. 30571034 and 30570424); the National Science Foundation of Heilongjiang Province (Grant Nos. D2007-48); Master Innovation Funds of Harbin Medical University (HCXS 2008010).

Conflict of Interest: none declared.

REFERENCES

Abeer *et al*. (2006) Immunological changes in patients with mania: changes in cell mediated immunity in a sample from Egyptian patients. *Egypt J. Immunol.*, **13**, 79–85.

Backes,C. *et al*. (2007) GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Res.*, **35**, W186–W192.

Bezchlibnyk,Y. and Young,L.T. (2002) The neurobiology of bipolar disorder: focus on signal transduction pathways and the regulation of gene expression. *Can. J. Psychiatry*, **47**, 135–148.

Blair,I.P. *et al*. (2006) Positional cloning, association analysis and expression studies provide convergent evidence that the cadherin gene FAT contains a bipolar disorder susceptibility allele. *Mol. Psychiatry*, **11**, 372–383.

Fukui,M. *et al*. (2008) Serum prostate-specific antigen levels in men with type 2 diabetes. *Diabetes Care*, **31**, 930–931.

Gould,T.D. and Manji,H.K. (2002) The Wnt signaling pathway in bipolar disorder. *Neuroscientist*, **8**, 497–511.

Imig,J.D. (2006) Eicosanoids and renal vascular function in diseases. *Clin. Sci.*, **111**, 21–34.

Kanehisa,M. (2002) The KEGG database. *Novartis Found. Symp.*, **247**, 91–101; discussion 101–103, 119–128, 244–152.

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

Kim,H. *et al*. (2005) A review of the possible relevance of inositol and the phosphatidylinositol second messenger system (PI-cycle) to psychiatric disorders–focus on magnetic resonance spectroscopy (MRS) studies. *Hum. Psychopharmacol.*, **20**, 309–326.

Lyoo,I.K. *et al*. (2003) Oral choline decreases brain purine levels in lithium-treated subjects with rapid-cycling bipolar disorder: a double-blind trial using proton and lithium magnetic resonance spectroscopy. *Bipolar Disord.*, **5**, 300–306.

Mocellin,S. *et al*. (2004) Molecular oncology in the post-genomic era: the challenge of proteomics. *Trends Mol. Med.*, **10**, 24–32.

Mohamad,M.J. *et al*. (2006) Serum sex hormones in premenopausal women with coronary heart disease. *Neuro Endocrinol. Lett.*, **27**, 758–762.

Montezinho,L.P. *et al*. (2007) Effects of mood stabilizers on the inhibition of adenylate cyclase via dopamine D(2)-like receptors. *Bipolar Disord.*, **9**, 290–297.

Ogata,H. *et al*. (2000) A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res.*, **28**, 4021–4028.

Rosenberg,G. (2007) The mechanisms of action of valproate in neuropsychiatric disorders: can we see the forest for the trees? *Cell. Mol. Life Sci.*, **64**, 2090–2103.

Sjoholt,G. *et al*. (2004) Examination of IMPA1 and IMPA2 genes in manic-depressive patients: association between IMPA2 promoter polymorphisms and bipolar disorder. *Mol. Psychiatry*, **9**, 621–629.

Tian,M. *et al*. (2008) Confidence intervals for the risk ratio under inverse sampling. *Stat. Med.*, **27**, 3301–3324.

Wang,K. *et al*. (2007) Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.*, **81**, 1278–1283.

WTCCC (2007) Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature*, **447**, 661–678.