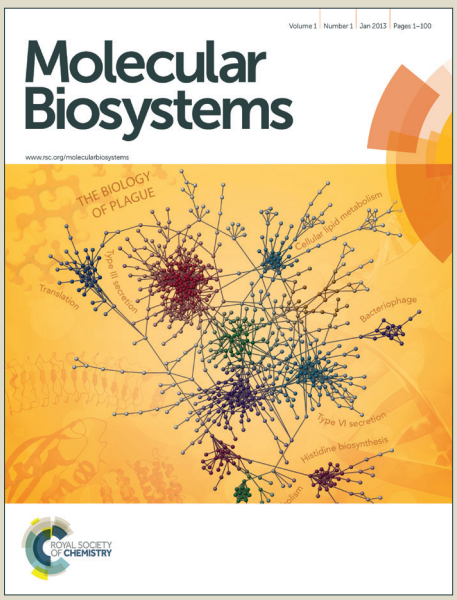
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Prioritization of Rheumatoid Arthritis Risk Subpathways Based on Global Immune Subpathway Interaction Network and Random Walk Strategy

Wenhua Lv[1, †], Qiuyu Wang[2, †], He Chen[3, †], Yongshuai Jiang[1, †], Jiajia Zheng[1], Miao Shi[1], Yanjun Xu[1], Junwei Han[1], Chunquan Li[4,*], Ruijie Zhang[1,*]

[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150086, China, [2] School of Nursing, Daqing Campus Harbin Medical University, Daqing 163319, China, [3] Depatment of Pathology, Harbin Medical University, Harbin 150086, China, [4] School of Medical Informatics, Daqing Campus Harbin Medical University, Harbin 150086, China

*To whom correspondence should be addressed: zhangruijie2013@gmail.com (Ruijie Zhang) Tel: +86045186650721-106, Fax: +86045186615922.and lcqbio@163.com(Chunquan Li) Tel: +86 04598153035; Fax: +8604598153035
† Joint First Authors.

Abstract

The initiation and development of rheumatoid arthritis (RA) is closely related to mutual dysfunction of multiple pathways. Furthermore, some similar molecular mechanisms are shared between RA and other immune diseases. Therefore it is vital to reveal the molecular mechanism of RA through searching for subpathways of immune diseases and investigating the crosstalk effect among subpathways. Here we exploited an integrated approach combining both construction of subpathway-subpathway interaction network and random walk strategy to prioritize RA risk subpathways. Our method can be divided into three parts: (1) acquisition of risk genes and identification of risk subpathways of 85 immune diseases by using Subpathway-LDS method; (2) construction of a global immune subpathway interaction (GISI) network with subpathways identified by Subpathway-LDS; (3) optimization of RA risk subpathways by random walk strategy based on GISI network. The results showed that our method could effectively identify RA risk subpathways, such as MAPK signaling pathway, prostate cancer pathway and chemokine signaling pathway. The integrated strategy considering crosstalk between immune subpathways significantly improved the effect of risk subpathway identification. With the development of GWAS, our method will provide insight into exploring molecular mechanisms of immune diseases and might be a promising approach for studying other diseases.

Key words: rheumatoid arthritis; immune disease; subpathway; interactive network; random walk

Introduction

The initiation and progression of diseases are closely related to the dysfunction of pathways. It is a huge challenge in functional genomics era to investigate molecular mechanisms of complex diseases including immune disease [1, 2]. However, molecules in cells seldom function in isolation, but interact with each other thus forming complex cellular pathways of metabolic, regulatory, or protein complexes to perform biological functions[3, 4]. Each biological pathway interacting with others is a member of complex biological network. Immune system diseases which are complex diseases are characterized by imbalance of immune regulation and dysfunction of the immune system, thus affecting immune response. Therefore, the independent analysis of one single gene or molecule fails to interpret the mechanism of these diseases. Studies about pathways of immune diseases and the interaction among different pathways will provide new insights for deeply understanding immune diseases.

The existing data related with immune diseases including (1) gene and SNP data and (2) pathway data help to systematically investigate immune disorders. There are many databases that store a large number of immune disease genes verified by experiments, such as the Online Mendelian Inheritance In Man (OMIM) database containing single gene study results for various immune diseases and the Genetic Association Database (GAD) storing lots of risk genes of immune disease proved by low-throughput experiments and information of other complex diseases. In addition, with the increasing number of high-throughput studies, SNPs identified by genome wide association studies (GWAS) become available for research, such as data of Wellcome Trust Case Control Consortium(WTCCC). In recent years, many biological pathways represented with different forms mainly including protein interaction, metabolic pathways and signaling pathways have been revealed and stored in pathway databases. A protein interaction database generally stores interaction networks of proteins. The most popular protein databases are STRING, BIND and HPRD. Despite the wide employment in studying human diseases, the protein interaction data is undirected which is inconsistent with the real molecule interaction in pathways and

has a great impact on the accuracy of result. Fortunately, the tricky problem can be settled by metabolic pathways and signaling pathways. Therefore, we chose KEGG, a commonly used pathway database with real-time updates containing both metabolic pathways and signaling pathways as the source of pathways.

It is important to develop methods for identifying both risk pathways and local regions of pathways. Therefore, many approaches have been developed for searching risk pathways closely related with diseases[5-10]. The most widely used method is based on enrichment analysis theory which computes the number of risk genes and genes from background annotated to a certain pathway respectively then calculate the P value used for estimating the significant degree that risk genes annotated to this pathway compared with random situation. For instance, PathMAPA[11]and DAVID[12] are based on Fisher exact test and adjusted Fisher test respectively, while KOBAS[13] is based on cumulative hypergeometric test. The major defect of these scale-based methods is that they are dependent on the choice of threshold which has serious effect on the result. Gene set enrichment analysis (GSEA) [14]is a popular scale-free method that is independent of threshold particularly initially designed for gene expression profiles. All genes in the list are ranked based on the correlation between their expression level and the phenotype by choosing a suitable measure. Given an interesting set of genes S, such as genes in a pathway, GSEA aims to determine whether genes in S are randomly distributed throughout L or primarily found at one end of the list. Compared with scale-based methods, GSEA method performs better in identification of risk pathways owing to the contribution of both risk genes and non-risk genes classified by a certain measure used for evaluating the association degree with diseases. An impact analysis method developed by Draghici[15] includes the classical cumulative hypergeometric statistics but also takes into account other crucial factors such as the degree of genes' expression changes, the type and positions in the given pathways and interactions between genes, etc. Given that the occurrence and development of cancers is closely related to multiple pathways and interactions among them rather than a single pathway, Pham et al. [16]proposed a latent pathway

identification analysis (LPIA) which addressed crosstalk between different pathways and constructed a pathway-pathway interaction network. In recent years, with the development of high-throughput experimental technologies and biology databases, several pathway analysis methods combined with GWAS came about. For instance, Wang et al[17] developed a pathway-based approach for analysis of GWA studies, which modified the classical GSEA algorithm to prioritize risk pathways. They calculated a test statistic value such as χ^2 statistic in case-control research for each SNP , took the maximum statistic for all SNPs located near a gene (<500kb) as the significance of the gene and sorted the genes according to their statistics. Then a enrichment score for a particular gene set was computed analogous to GSEA method. Subsequently, some other methods[18, 19] based on or modified by this approach for analyzing GWA data appeared, these pathway-based approaches typically ranked all the genes by their significance statistic value or p value and then decided whether a particular group of genes was enriched at one end of the ranked list more than that of random situation.

Although it works well for these methods in finding risk pathways associated with complex diseases, these methods can only identify entire pathways; however, even the smallest pathway still contains at least tens to hundreds of genes. Even though these pathways are indeed associated with diseases, they are not precise enough. Therefore, it is essential to improve accuracy level by developing novel methods for searching for subpathways that are the local regions of an entire pathway. Many studies suggest that the initiation and progression of diseases are highly associated with abnormalities in the mutual dysfunction of multiple pathways [2, 3]. Hence identification of risk subpathways of immune diseases achieves more precise pathway result owing to the larger proportion of key genes contained in subpathways than that of entire pathway, and the subpathway-subpathway network performs better in reflecting interactions among different pathways.

Here we exploited an integrated approach to prioritize risk subpahtways of RA combining both identification of subpathways of immune diseases and optimization of RA risk pathways by using random walk strategy based on subpathway-subpathway

interaction network. Our method can be divided into the following three parts: (1) identification of risk subpathways of 85 immune diseases including rheumatoid arthritis (RA) by using Subpathway-LDS method previously developed by our group; (2) construction of global immune subpathway interaction (GISI) network; (3) optimization of RA risk subpathways by random walk strategy based on GISI network. The advantage of this method is that subpathway-subpathway network of immune diseases are treated as background and used to optimize RA subpathways found by Subpathway-LDS method. RA, as we all know, is one of autoimmune diseases belonging to immune diseases according to Medical Subject Headings (MeSH) category, sharing similar molecular mechanisms with other autoimmune diseases or other immune diseases. The evaluation of risk level of RA subpathways will be improved by taking full advantage of interaction information in network.

Results

Significant subpathways identifiedy by Subpathway-LDS

3909 risk genes were achieved after mapping risk SNPs to genome and 236 subpathways associated with RA were identified by Subpathway-LDS method. The number of risk subpathways reached 44 when the P-value threshold of cumulative hypergeometric test was 0.01, and increased to 85 while the threshold was 0.05. In addition, 704 risk subpathways of immune diseases were found by Subpathway-LDS.

In order to evaluate the power of Subpathway-LDS, we observed the top ten subpathways of whole subpathway list in ascending order of P-values (Table1). For instance, the focal adhesion pathway (p = 6.94E-09) ranking first at the whole list contained 173 genes, including 64 genes annotated by risk SNPs. Research showed that focal adhesion pathway could be a target pathway for treating RA[20]. Genes in focal adhesion pathway functioned as the bridge between endothelial cells and extracellular matrix. Dysfunction of this pathway is closely related to the pathogenesis of RA[21]. ErbB, a member of epidermal growth factor receptor(EGFR), is a cell-membrane receptor coded by oncogene erbB-2. Immunoreactivity for ErbB2 was

found by Hallbeck et al. [22] in synovial membrane especially for RA patients. The increased expression of ErbB in synovial fluid of RA patients may lead to an abnormal growth pattern, indicating the ErbB signaling pathway was related to RA. In addition, we found that there are three ECM-receptor interaction subpathways including RA-path:04512_9, RA-path:04512_6 nd RA-path:04512_10 in the top ten risk subpathways of RA. Although there are little evidence suggesting the direct relationship between ECM-receptor interaction and RA, ECM-receptor interaction can be involved in immune regulation and further affect immune response. The extracellular matrix (ECM) is composed of both structural and functional macromolecules and plays a vital role in tissue and organ morphogenesis and in the maintenance of cell and tissue structure and function. Interactions of Cell–cell and cell–ECM are critical in various developmental processes, such as proliferation, differentiation, and migration of cells. ECM proteins influence cellular functions thus forming a complex feedback mechanism. Versican, an extracellular matrix (ECM) proteoglycan interacts with cells by binding to integrin receptors or non-integrin receptors and to other ECM components associated with the cell surface. By binding to hyaluronan, versican influences phenotypes of T lymphocyte and partly affects the ability of synthesizing and secreting cytokines that influence the immune response[23]. Versican is an important ECM molecule that is critical for inflammation and might be a potential therapeutic target for immune diseases. These results show that the risk subpathways identified by Subpathway-LDS are mainly associated with RA or other immune diseases and inflammatory diseases.

Global immune subpathway interaction (GISI) network

We constructed a global immune subpathway interaction (GISI) network, in which nodes were risk subpathways of immune diseases and edges represented interaction among subpathways and P-value of Fisher exact test was treated as the weight of edges in the network (Figure 1A). There are 701 nodes and 30228 edges in the network. The average clustering coefficient is 0.822 (distribution of clustering coefficient see Fig 1C), indicating that nodes in GISI network are highly clustered.

Although the diameter of GISI network is 11, the distribution of shortest path length (Figure 1D) shows that distance between most nodes is short, suggesting that most nodes in the network are connected directly. Then we found an interesting result that the node degree distribution did not follow power-law distribution and the majority of nodes in the network had high degree, which differed from most of the biology network (Figure 1A). The average degree of the network reaches 86 (distribution of degree see Figure 1B), which means a subpathway node in the network averagely connects 86 other nodes, indicating tight interaction. Nodes connected with a certain node were divided into two classes, one was the nodes belonging to the same disease with the node we considered, and others were those belonging to the other diseases. For each of the 701 nodes, we calculated the number of two kinds of nodes and the ratio of them, represented by Nps, Npd and Rpd (Rpd= Npd/(Npd+ Nps)) and found that there were 86.9% (609/701) whose Rpd>=0.80. We further analyzed the global crosstalk between a certain disease and all the other immune diseases in our network. The interaction edges in the network connecting two nodes fell into two types which referred to edges connecting two nodes pertaining to the same disease and different diseases respectively. For each of the 62 disease, we computed the number of two kinds of interactions and the percentage of the second one, denoted by Nds, Ndd and Rdd (Ndd/ (Ndd+ Nds)) respectively. Rdd ranges from 0.932 to 1, showing that for each of the disease in the network, the proportion of interaction with other different diseases is above 90%.

The above analysis demonstrates that the average degree of global immune interaction network is high which means that there exists closely tight interaction between subpathways. For subpathways that belong to a same disease, if they share some common genes, they are likely to be merged as a single subpathway, while for different diseases which belong to the same category such as immune diseases, the subpathways that have some common genes will be regarded as different subpathways and kept for further analysis. Therefore, the interaction degree between each disease's subpathways and all the other diseases' subpathways is higher than that

of the same disease. Besides, the higher interaction degree between subpathways of RA and the other subpathways of other immune diseases also indicates that RA share similar genetic mechanism with other immune diseases, thus our network could provide abundant information for the identification of risk subpathway of RA. Compared with the conventional method, the subpathway interaction network of immune diseases takes much more factors into account, thus guaranteeing better results when applied to identification of RA risk subpathways.

A new large network including 878 nodes and 40067 interactive edges came about after adding relevant subpathways of RA to the immune network, in which blue nodes represent RA subpathways, gray nodes represent subpathways of other immune diseases, and red nodes denote the top ten RA subpathways with high random walk scores (Figure 2).The reason why they score high is that there is strong crosstalk effect between red nodes and their neighbors.

RA risk subpathways optimized by random walk strategy

For each of the 236 RA subpathways, the P-value and random walk score were calculated by Subpathway-LDS and random walk algorithm respectively. Table 2 lists the top ten RA subpathways with high random walk values. To evaluate the effectiveness of immune network in identifying RA risk subpathways, we observed the relationship between P-values and random walk scores (Figure 3). 85 significant RA subpathways ($p<0.05$) were found using Subpathway-LDS method, of which 87.1% (74/85) had scores above average score of all the RA subpathways. The result indicates that majority of the risk subpathways found by Subpathway-LDS method score higher than those non-risk subpathways while using our optimization method. On account of the high reliability of Subpathway-LDS, the results optimized by random walk strategy are highly credible. Mitogen-activated protein kinase (MAPK) signaling pathway and prostate cancer pathway are the most typical examples. The p values of MAPK signaling pathway and prostate cancer pathway are 2.36E-06 and 0.0038 respectively, indicating that both of them might be related with RA. In addition to the low p values, they got high random walk scores and ranked top ten in

the whole list, proving that our method performs well in prioritizing risk subpathways. MAPK signaling pathway, known as a target pathway for treating RA, was reported to be associated with pathology of RA[24]. Interleukin-1β (IL-1β), one of the inflammatory cytokines, plays an important role in the development of RA, meanwhile it is an inhibitor for MAPK signaling pathway[25]. Inhibition of MAPK signaling pathway prevent fibroblast-like synoviocyte (FLS) cells from growing. A research showed that about 1/3 genes of FLS were regulated by MAPK signaling pathway , indicating the possibilities for being treated as target pathway of RA drugs. Serum concentrations of Chromogranin A (CgA) that is a crucial neuroendocrine tumor marker can reflect the activity of neuroendocrine and evaluate the progress of prostate tumor. The treatment for RA increases the serum concentrations of CgA, which illustrates that RA is related with prostate cancer [26]. A recent study showed that Dickkopt-1（DKK-1）, an inhibitor for Wnt pathway, which was critical for prostate cancer bone metastasis, may be involved in the remodeling process of RA[27]. SR31747A currently being evaluated in phase IIb clinical effectiveness trials for prostate cancer treatment is an agent with immunomodulatory and antiproliferative activities. The molecule can prevent lymphocytes of human and mouse from proliferating, modulate the expression of pro-inflammatory and anti-inflammatory cytokines, and was shown to protect animals against acute and chronic inflammatory conditions, such as RA[28].

It is worth noting that though some subpathways do not meet the threshold of P-value, they get high random walk score, such as Jak-STAT signaling pathway (P=0.179, see Figure 4), purine metabolism pathway (P=0.071), natural killer cell mediated cytotoxicity pathway (p=0.0564, see Figure 5) and chemokine signaling pathway (P=0.105). As a typical subpathway of RA, Jak-STAT signaling pathway has been widely used to study the disease and treated as a target pathway of treatment drugs. Several cytokines such as Tumor Necrosis Factor (TNF), and certain Interleukin are critical for the development of RA. A study reported that Interleukin-2 worked by activating JAK-STAT signaling pathway and played crucial roles in leukomonocyte development, thus further affecting the immune response after organ

transplantation[29]. Some researchers proved the relationship between Jak-STAT signaling pathway and RA because the JAK inhibitors functions well in the clinical treatment of RA. Given the importance of Jak-STAT pathway in the pathogenesis of RA, many researches concentrated on the biological agents targeting JAKs and demonstrated that JAK inhibitors would serve as the most promising new agents for treatment of RA[30]. One of the most advanced JAK inhibitors in treatment of RA is CP-690550(tasocitinib and tofacitinib), Migita et al conducted an experiment to assess the effects of CP-690550 on JAK inhibition, and came to the conclusion that inhibition of these proinflammatory signaling pathways by CP690550 could be important in the treatment of RA[31]. Similarly, purine metabolism pathway was involved in immune system[32] because of adenosine dehydrogenase. Taking methotrexate also resulted in reduced activity of purinase [33] thus proving purine metabolism pathway was closely related with RA. Natural killer (NK) cells are lymphocytes of the innate immune system involved in the early defenses against allogeneic cells, as well as autologous cells that undergo various forms of stress, including infection with bacteria, viruses, or parasites or malignant transformation[34]. NK cells work in two primary ways: one is detecting and vitiating transformed cells and cells infected by virus, the other is secreting diverse cytokines that are critical for the innate and adaptive immune responses[35]. The activation of NK cells is regulated by both activating receptors and inhibitory receptors. DAP10 and NKG2D form a complex thus activating the immunological competence of NK cells. Although they are not risk genes of RA, DAP10 and NKG2D simultaneously appear in the NK cell mediated cytotoxicity subpathway (Figure 5) indicating the possibilities of relationship between this subpathway and RA. These results show that JAK-STAT signaling pathway, purine metabolism pathway and NK cell mediated cytotoxicity pathway are highly associated with RA and the random walk strategy works well in optimizing potential risk subpathways of RA.

Interestingly, besides those pathways that have been proved to be related with RA or other immune diseases, we also found some subpathways which had never been

reported to be associated with RA, such as chemokine signaling pathway. Despite the large p value (p=0.1), chemokine signaling pathway ranked first at the score list. Few researches reported the direct relationship between RA and chemokine signaling pathway, but a study showed that chemokine signaling pathway was involved in activation of CCL2[36]. Some researchers analyzed gene expression in synovium using collagen-induced arthritis (CIA) rat model and found that gene expression level of CCL2 increased in CIA. The expression of CCL2 decreased after treating CIA rat with low intensity laser radiation. Besides, our research team also performed a genome-wide haplotype association analysis and gene prioritization for identifying risk locus of RA, which ranked all the candidate RA risk genes based on both structural similarity and functional similarity, then found that 4 CCL genes appeared at the top ten risk gene list of RA[37]. The result shows that this pathway may be associated with RA in animal model and genes of CCL family are closely related to RA, indicating that the pathway is possibly believed to be a potential novel risk subpathway for RA.

To deeply interpret the reason why those RA subpathways such as MAPK signaling pathway (RA-path:04010_1), Jak-STAT signaling pathway (RA-path:04630_1) and chemokine signaling pathway (RA-path:04062_1) score higher than other subpathways, we constructed sub-networks with first neighbors of these three risk subpathways. These three sub-networks contained 240, 164, 293 nodes and 17119, 5811, 19636 edges respectively. We further classified the subpathway nodes in the subnets connected with key nodes including RA-path:04010_1, RA-path:04630_1 and RA-path:04062_1 by diseases and calculated the number of subpathways for each disease. Then we found that the number of subpathways of some diseases is much larger than that of other diseases, such as arthritis, systemic lupus erythematosus, type I dibetes and Crohn's Disease, which mostly belonged to autoimmune diseases or inflammatory disease (Table 3-5). Owing to the same or similar molecular mechanism shared with other autoimmune diseases and inflammatory diseases, RA related risk subpathways could get much information from subpathways of those diseases

mentioned above and then get much higher random walk scores than other subpathways. This analysis illustrates that subpathways of different immune diseases are closely connected by the GISI network which does provide useful information for identifying and optimizing risk subpathways.

In conclusion, it is quite easy to find that to some extent the random walk stratedy based on GISI network is superior to those conventional methods based on hypergeometric test. Owing to the additional important information provided by subpathways of other immune diseases in the network, known and novel risk subpathways can be identified by our method. Moreover, some false positive results are filtered out by using our method, for example, Gastric Acid Secretion and some cancer pathways that are apparently unrelated with RA ranked after their original positions while using our novel method.

Methods

Firstly, we obtained risk SNPs of RA from WTCCC, determined its risk genes according to the positions of risk SNPs and obtained risk genes of other immune diseases from GAD. Secondly, we utilized subpathway-LDS algorithm[5] previously developed by our group to identify risk subpathways of the immune diseases including RA. Furthermore, a GISI network was constructed using risk immune subpathways identified through Subpathway-LDS method. Finally, RA risk subpathways identified by Subpathway-LDS was optimized by random walk strategy, treating GISI network as background.

Risk SNPs and genes of RA

WTCCC is a consortium aiming to conduct genome wide association study and collects SNP data associated with several diseases, such as rheumatoid arthritis, diabetes mellitus and coronary artery disease. Association studies were performed for a certain disease and P values of all SNPs representing significant level were achieved. We obtained 10890 RA risk SNPs ($P<0.05$) from WTCCC and 3909 risk genes by mapping SNPs to genes according to physical locations on the chromosome. These

risk genes were treated as input data of Subpathway-LDS method for identifying risk subpathways of RA.

Risk genes of immune diseases

The GAD is the NIH supported public repository of human genetic association studies of complex diseases, which contains the complete known gene-phenotype associations and includes data of non-mendelian common complex diseases. We used phenotype-gene data approved by experiment and reference in GAD for constructing network. The whole phenotype-gene relationships were downloaded, and then gene ID was unitized with Entrez ID. After filtering, only 2827 unique phenotype-gene pairs of 85 immune diseases were left.

Identification of risk subpathways of immune diseases based on Subpathway-LDS method

We used the Subpathway-LDS method, of which "LDS" means lenient distance similarity, to identify the risk subpathways of rheumatoid arthritis (RA) and other immune diseases. First, the KEGG pathways were converted to directed graphs; second, signatures nodes were located in the directed graphs according to risk genes; third, subpathways in which the number of nodes was not less than s were identified when the shortest path length between two signature nodes was shorter than n+1; fourth, the significance of subpathways was evaluated by hypergeometric tests. The subpathways we identified were used for constructing the GISI network and further optimizing the risk subpathways of RA. The detailed algorithms were specified as follows:

(Ⅰ) Convert each pathway to a directed graph

The KGML files from KEGG database were downloaded to obtain the relationships of genes in the corresponding pathways. For metabolic pathways and unmetabolic pathways, two methods were used during converting pathways to directed graphs. Specifically, metabolic pathways used enzymes as node sets of graphs, and generated edge sets from biochemical reactions. Two nodes in a directed graph were connected by an edge if two enzymes were involved in a continuous biochemical reaction,

indicating that they shared common metabolites. An arrow was pointed from one enzyme to the other if the product of first enzyme was the substrate of the second one and vice versa. Two directed edges showed up when the reactions were reversible. Two enzyme nodes will not be connected if the common metabolite they share is the substrate or production in first reaction and the same in second reaction. It is a remarkable fact that one enzyme node may appear at two or more locations in the pathway, involve different biochemical reactions and therefore perform different functions. Different from the most majority of other methods, which usually merged the enzymes with multiple locations in the pathway, our method kept them as what they were. In this way, the directed graphs we got are more similar to the original pathways, helping to improve the accuracy of risk subpathways identified by Subpathway-LDS method. Directed graphs of unmetabolic pathways used all proteins in the pathways as their nodes, and the edges were determined by the interaction relationships of genes in the pathways. For instance, a directed edge would be pointed from a transcription factor to the corresponding target gene if the transcription factor activates the target gene. The progress of converting pathways to directed graphs was performed by using the iSubpathwayMiner[38], which was a tool designed for analyzing pathways.

(Ⅱ) Locate signature nodes within pathways according to risk SNPs

First, 3903 risk genes of RA were obtained the projection of risk SNPs on the basis of their physical positions on the chromosomes. Then risk genes were mapped to converted directed graphs and those annotated nodes in the graphs were named by signature nodes. For the other 85 immune diseases in GAD, signature nodes were obtained by directly mapping risk genes to the pathway graphs.

(Ⅲ) Identify subpathway regions by using LDS strategy

For each pathway containing signature nodes, we computed the shortest path length of any two signature nodes in the given directed pathway graph. The two signature nodes and other non-signature nodes would be added to a same node set, when the shortest path length between two signature nodes was shorter than n+1, in which the

parameter n indicated the maximum permitted number of non-signature nodes at the shortest path between two signature nodes. We then extracted the corresponding subgraphs in the directed pathway graph according to the node set we got and filtered subpathways of which the number of nodes was less than s. The parameter s was to make sure that there were enough nodes in the subgraphs because a subgragh with small amount of nodes could not form a biological subpathway. On the other hand, flexibility was introduced to this subpathway strategy by adjusting the parameter n. A smaller value of n means that only those nodes meeting stricter distance similarities will be added to the corresponding subpathway, and thus fewer subpathways will be identified and the number of non-signature nodes will reduce compared with larger values of n. Here we set the parameters based on the results of a previous research performed by our research team[5]. In that study, we computed the shortest distance between each disease node and its nearest disease node and found that the distance was <5 for 85% disease nodes. Therefore, we set n=5 and s=3. The lenient distance similarity in our method means both signature and non-signature nodes may be added to a subpathway so long as they meet the criteria of shortest path length. Nodes with higher topology centrality, such as degree and betweenness will be more likely to show up in the final risk subpathways.

(Ⅳ)Evaluate the statistical significance of subpathways

For each subpathway identified by Subpathway-LDS, hypergeometric test, one of the most commonly used statistical approaches, was used to calculate the statistical significance.The number of risk genes (r) annotated to the subpathway was counted and then the probability value that the number of risk genes randomly annotated to the subpathway is larger than r is calculated, in which a smaller P-value indicates that the risk subpathway is more closely associated with RA or other immune diseases compared to higher P-value. The equation Ⅰ represents an example calculation of the statistical significance of a subpathway, in which m denotes the number of genes in human genome serving as the background gene set, n denotes the number of risk genes to which risk SNPs map, t and r are the number of genes and risk genes in each

pathway, respectively.

$$P = 1 - \sum_{x=0}^{r-1} \frac{\binom{t}{x}\binom{m-t}{n-x}}{\binom{m}{n}} \qquad (\text{I})$$

Construction of global immune subpathway interaction network

In consideration of the similar molecular mechanisms shared by immune diseases and the crosstalk effect between different subpathways, we constructed a global immune subpathway interaction (GISI) network as background for prioritizing risk subpathways of RA. Firstly we treated risk genes of immune diseases obtained from GAD as the input data of Subpathway-LDS and got risk subpthways of 85 immune diseases. Then we built the GISI network using these risk subpathways as nodes in the network. In the GISI network, nodes are those risk subpathways identified by subpathway-LDS approach and edges connecting two nodes represent crosstalk between two risk subpathways. Two subpathways will be connected with an edge if they share common genes. For each subpathway pair which had common genes, Fisher test was performed and the P-value was considered as the weight of edge, in which a smaller P-value meant more tight crosstalk. For two different subpathways, we respectively counted the number of common genes (represented by a), the number of genes that only appeared in subpathway 1(represented by b), the number of genes that only appeared in subpathway 2 (represented by c) and the difference between the number of backgroud genes and the total number genes in both subpathway 1 and subpathway 2 (represented by d). The above 4 numbers were used to calculate the p-value according to the formula of Fisher exact test (equation Ⅱ), where N is the sum of a, b, c and d.

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!} \qquad (\text{II})$$

It is worth noting that the threshold setting step was omitted and all the subpathways identified by Subpathway-LDS were used as nodes of the network owing to the high reliability of the risk genes of immune diseases.

Prioritization of RA risk subpathways based on random walk strategy

RA risk subpathways identified by Suppathway-LDS were optimized by random walk strategy [39] in which the GISI network functions as the background. The random walk strategy based on GISI network takes full advantage of information iteratively transferred in the network. For each of the RA subpathways, we calculated a stable state score to reflect the importance of the risk subpathway. The basic idea of random walk strategy is to give corresponding scores for the nodes of network based on their topological importance. Nodes located at more important positions of topological structure are prone to get higher random walk scores when reaching the steady state by multiple iterations. First, equal initial values are assigned for all the nodes in the network, for instance the initial value of each node will be 1/n when the total number of nodes in the network is n, and then iteration steps are repeated according to the following equation Ⅲ.

$$P^{t+1} = (1-\gamma)WP^{t} + \gamma P^{0} \qquad (Ⅲ)$$

In this equation, P^0 denotes the initial value vector of all nodes in the network, P^t and P^{t+1} denotes value vectors at time t and t+1 respectively, W is an adjacency matrix with n×n elements determined by the number of common genes shared by two subpathways, and γis a constant and was set to 0.7 in our study. The iteration repeats until the D-value between P^t and P^{t+1} is lower than a certain threshold (1e-10), suggesting that network is reaching the stable state. Since the score value of a certain node at time t+1 equals to the sum of scores transferred from other nodes at time t and its own initial value at time t^0, a RA subpathway node will score higher suggesting higher risk degree if it shares more genes and nodes connected with this subpathway

simultaneously share more genes with their neighbors. Since the score value of a subpathway at time t+1 is calculated through adding the score values transferred from other subpathways at time t and the its own initial value at time t^0, a RA subpathway will get higher score if it shares more genes with other subpathways and nodes connected with this subpathway simultaneously share more genes with their neighbors, suggesting that this RA subpathway is in an important position in the global immune network and tend to be a risk subpathway of RA. Sorted by score values in ascending order, subpathways at the top of list are the optimized risk RA subpathways.

Discussion

The initiation and development of RA which is a complex disease are closely related to mutual dysfunction of multiple pathways. Multiple pathways constitute a complex biological network by pathway-pathway crosstalk. Therefore the independent analysis of one single gene or molecule fails to interpret the mechanism of these diseases. Here we exploited an integrated approach to prioritize RA risk subpahtways combining both construction of subpathway-subpathway interaction network and random walk strategy.

Compared with conventional pathway identification method based on enrichment analysis, Subpathway-LDS method searches for subpathways instead of entire pathways so that the accuracy of result can be largely increased. On the other hand, flexibility was introduced to this subpathway strategy by adjusting the parameter n (the maximum permitted number of non-signature nodes at the shortest path between two signature nodes) and s (the minimum permitted number of nodes in a subpathway). Here we set n=5 and s=3 aiming to achieve both comprehensive and accurate results.

Risk subpathways of immune diseases were identified by Subpathway-LDS and then a GISI network of immune diseases was constructed serving as a background to optimize the RA risk subpathways identified by Subpathway-LDS. In this study, we fully considered the crosstalk effect between a subpathway and another because RA

shared similar molecular mechanisms with other immune diseases. We chose risk genes of immune diseases in GAD rather than risk genes obtained from WTCCC data as the input data of Subpathway-LDS method for identifying risk subpathways because GAD contains comprehensive risk genes of much more immune diseases and results in GAD are validated by low-throughput experiments. By far GAD is one of the most reliable disease association database which not only includes abundant disease-gene information and but also provides corresponding literature support. In this study, we used the WTCCC data (genome wide association study) to identify the risk RA subpathways. Therefore, it is more appropriate to construct background immune subpathway interaction network with GAD then other databases. KEGG was chosen as the source of pathway data because KEGG is the most commonly used and pathway database which provides free available data. The pathway data of KEGG are well organized and directed, which is appropriate for identifying risk subpathways with the subpathway-LDS approach.

The topological property analysis demonstrates that the average degree of global immune interaction network is high which means that majority of subpathways in the network closely interact with others suggesting that subpathways of other immune disease do provide useful information for identification and optimization of RA risk subpathways. Random walk strategy based on subpathway network was utilized to prioritize risk subpathways of RA. A stable state score was calculated to reflect the importance of the nodes in the network by iteratively transferring information. For the GISI network, the nodes with topological importance are prone to be the risk subpathways of immune diseases. The random walk algorithm is often used to identify significant nodes including genes (proteins) and pathways which are composed of genes (proteins) of the network. In recent years, the random walk strategy has been widely used in the prediction of risk genes of diseases. For instance, Ko¨hler et al.[40] applied the random walk approach for prioritization of candidate genes of diseases and achieved better results than previous methods. Besides, our research team has applied the random walk algorithm to predict of survival time of

cancer patients[41], to prioritize candidate disease metabolites[42], to identify risk pathways[43]or the cross-talk among different pathways[44].

Each RA subpathway obtained by Subpathway-LDS gets a score for evaluating risk level after being optimized by random walk strategy based on subpathway network of immune diseases. The risk RA subpathways optimized by random walk strategy fall into 4 types: (1) majority of the RA risk subpathways found by Subpathway-LDS method including MAPK signaling pathway and prostate cancer pathway score higher than those non-risk subpathways while using our optimization method; (2) although some subpathways do not meet the threshold of P-value, they get high rank while exploiting random walk strategy, such as Jak-STAT signaling pathway (P=0.179), purine metabolism pathway (P=0.071) and chemokine signaling pathway (P=0.105); (3) chemokine signaling pathway tends to be a potential novel risk subpathway of RA; (4) some false positive results are filtered out by using our method such as Gastric Acid Secretion and some cancer pathways.

For the top ten RA risk sub-pathways optimized by random walk strategy based on GISI network, we compared the random walk scores and the corresponding P values of hypergeometric test (Table 2). We found that five of the ten RA subpathways got P values larger than 0.05 meaning that these 5 subpathways (Jak-STAT signaling pathway, purine metabolism pathway, natural killer cell mediated cytotoxicity pathway, chemokine signaling pathway and retinol metabolism pathway) could not be identified by hypergeometric test. However, some of these subpathways that do not meet threshold of P value have been reported to be associated with RA or immune response. For instance, Jak-STAT signaling pathway, a typical subpathway of RA, has been used for the treatment of RA. The purine metabolism pathway, natural killer cell mediated cytotoxicity pathway have also been proved to be associated with immune reponses indicating that they are possible related with RA. In addition, some subpathways that has never been reported to be associated with RA were also identified by the random walk strategy. For instance, the chemokine signaling pathway ranked first at the score list, but its p value was larger than 0.05. Few

researches reported the direct relationship between RA and chemokine signaling pathway, but through analyzing the previous results associated with chemokine signaling pathway, we infered it might be a potential risk subpathway of RA. In conclusion, through comparison with the traditional hypergeometric test, the random walk strategy performs well in the prioritization of RA risk subpathways.

Compared with conventional methods based on hypergeometric test, we took into account the crosstalk between subpathways and constructed a global immune network with risk subpathways identified through Subpathway-LDS method. Subpathways in the network interact with each other and provide useful information for other subpathways, so that the risk subpathways prioritized by our method might be more subtle and reliable. Due to the limited amount of disease-related gene data, the GISI network is still not perfect in identifying risk subpathways of immune diseases. However, with the development of GWAS, our method will provide novel insight into exploring molecular mechanisms of immune diseases and might be a promising approach for studying other diseases.

References

1. A.-L. Barabási, *New England Journal of Medicine*, 2007, **357**, 404-407.
2. A. Zelezniak, T. H. Pers, S. Soares, M. E. Patti and K. R. Patil, *PLoS computational biology*, 2010, **6**, e1000729.
3. D.-S. Lee, J. Park, K. Kay, N. Christakis, Z. Oltvai and A.-L. Barabási, *Proceedings of the National Academy of Sciences*, 2008, **105**, 9880-9885.
4. J. D. Han, *Cell Res*, 2008, **18**, 224-237.
5. C. Li, J. Han, Q. Yao, C. Zou, Y. Xu, C. Zhang, D. Shang, L. Zhou, Z. Sun, J. Li, Y. Zhang, H. Yang, X. Gao and X. Li, *Nucleic Acids Res*, 2013, **41**, e101.
6. A. V. Antonov, S. Dietmann and H. W. Mewes, *Genome Biol*, 2008, **9**, R179.
7. B. Mlecnik, M. Scheideler, H. Hackl, J. Hartler, F. Sanchez-Cabo and Z. Trajanoski, *Nucleic acids research*, 2005, **33**, W633-W637.
8. J. Xia, N. Psychogios, N. Young and D. S. Wishart, *Nucleic acids research*, 2009, **37**, W652-W660.
9. J. Xia and D. S. Wishart, *Nucleic acids research*, 2010, **38**, W71-W77.
10. J. Xia and D. S. Wishart, *Bioinformatics*, 2010, **26**, 2342-2344.
11. D. Pan, N. Sun, K.-H. Cheung, Z. Guan, L. Ma, M. Holford, X. Deng and H. Zhao, *BMC bioinformatics*, 2003, **4**, 56.
12. G. Dennis Jr, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane and R. A. Lempicki, *Genome biol*, 2003, **4**, P3.
13. J. Wu, X. Mao, T. Cai, J. Luo and L. Wei, *Nucleic acids research*, 2006, **34**,

W720-W724.
14. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub and E. S. Lander, *Proceedings of the National Academy of Sciences of the United States of America*, 2005, **102**, 15545-15550.
15. S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu and R. Romero, *Genome research*, 2007, **17**, 1537-1545.
16. L. Pham, L. Christadore, S. Schaus and E. D. Kolaczyk, *Proceedings of the National Academy of Sciences*, 2011, **108**, 13347-13352.
17. K. Wang, M. Li and M. Bucan, *Am J Hum Genet*, 2007, **81**, 1278-1283.
18. L. Chen, L. Zhang, Y. Zhao, L. Xu, Y. Shang, Q. Wang, W. Li, H. Wang and X. Li, *Bioinformatics*, 2009, **25**, 237-242.
19. S. E. Baranzini, N. W. Galwey, J. Wang, P. Khankhanian, R. Lindberg, D. Pelletier, W. Wu, B. M. Uitdehaag, L. Kappos, C. H. Polman, P. M. Matthews, S. L. Hauser, R. A. Gibson, J. R. Oksenberg and M. R. Barnes, *Hum Mol Genet*, 2009, **18**, 2078-2090.
20. K. Nakano, J. W. Whitaker, D. L. Boyle, W. Wang and G. S. Firestein, *Annals of the rheumatic diseases*, 2013, **72**, 110-117.
21. M. Connolly, D. Veale and U. Fearon, *Annals of the rheumatic diseases*, 2011, **70**, 1296-1303.
22. A.-L. Hallbeck, T. M. Walz, K. Briheim and Å. Wasteson, *Scandinavian journal of rheumatology*, 2005, **34**, 204-211.
23. T. N. Wight, I. Kang and M. J. Merrilees, *Matrix Biology*, 2014.
24. A. R. Clark and J. L. Dean, *Open Rheumatol J*, 2012, **6**, 209-219.
25. M. Shakibaei, A. Mobasheri and C. Buhrmann, *Genes Nutr*, 2011, **6**, 171-179.
26. C. Zer, G. Sachs and J. M. Shin, *Physiol Genomics*, 2007, **31**, 343-351.
27. D. Daoussis and A. P. Andonopoulos, *Semin Arthritis Rheum*, 2011, **41**, 170-177.
28. P. Casellas, S. Galiegue, B. Bourrie, J.-B. Ferrini, O. Jbilo and H. Vidal, *Anti-cancer drugs*, 2004, **15**, 113-118.
29. R. Vafadari, W. Weimar and C. C. Baan, *Clin Chim Acta*, 2012, **413**, 1398-1405.
30. K. Vaddi and M. Luchi, *Expert Opin Investig Drugs*, 2012, **21**, 961-973.
31. K. Migita, A. Komori, T. Torigoshi, Y. Maeda, Y. Izumi, Y. Jiuchi, T. Miyashita, M. Nakamura, S. Motokawa and H. Ishibashi, *Arthritis Res Ther*, 2011, **13**, R72.
32. Z. Zakeri, S. Izadi, A. Niazi, Z. Bari, S. Zendeboodi, M. Shakiba, M. Mashhadi, B. Narouie and M. Ghasemi-Rad, *Int J Clin Exp Med*, 2012, **5**, 195-200.
33. S. L. Morgan, R. A. Oster, J. Y. Lee, G. S. Alarcon and J. E. Baggott, *Arthritis Rheum*, 2004, **50**, 3104-3111.
34. E. Vivier, J. A. Nunès and F. Vély, *Science*, 2004, **306**, 1517-1519.
35. A. A. Maghazachi, *Pharmacological reviews*, 2005, **57**, 339-357.
36. L. Zhang, J. Zhao, N. Kuboyama and Y. Abiko, *Lasers Med Sci*, 2011, **26**, 707-717.
37. R. Zhang, P. Sun, Y. Jiang, Z. Chen, C. Huang and X. Zhang, *International journal of immunogenetics*, 2010, **37**, 273-278.
38. C. Li, X. Li, Y. Miao, Q. Wang, W. Jiang, C. Xu, J. Li, J. Han, F. Zhang, B. Gong

and L. Xu, *Nucleic Acids Res*, 2009, **37**, e131.
39. H. Tong, C. Faloutsos and J.-Y. Pan, 2006.
40. S. Kohler, S. Bauer, D. Horn and P. N. Robinson, *Am J Hum Genet*, 2008, **82**, 949-958.
41. W. Liu, Q. Wang, J. Zhao, C. Zhang, Y. Liu, J. Zhang, X. Bai, X. Li, H. Feng, M. Liao, W. Wang and C. Li, *Mol Biosyst*, 2015.
42. D. Shang, C. Li, Q. Yao, H. Yang, Y. Xu, J. Han, J. Li, F. Su, Y. Zhang, C. Zhang, D. Li and X. Li, *PLoS One*, 2014, **9**, e104934.
43. W. Liu, C. Li, Y. Xu, H. Yang, Q. Yao, J. Han, D. Shang, C. Zhang, F. Su, X. Li, Y. Xiao, F. Zhang and M. Dai, *Bioinformatics*, 2013, **29**, 2169-2177.
44. J. Han, C. Li, H. Yang, Y. Xu, C. Zhang, J. Ma, X. Shi, W. Liu, D. Shang, Q. Yao, Y. Zhang, F. Su, L. Feng and X. Li, *J R Soc Interface*, 2015, **12**, 20140937.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (Grant Nos.31200934, 81172842) and the Natural Science Foundation of Heilongjiang Province, China (Grant No.C201206). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.Conflict of Interest statement. The authors declare they have no conflict of interests.

Figure Legends

Figure 1. The global immune disease subpathway crosstalk network and fundamental topological properties. (A) The subpathway crosstalk network of 85 immune diseases. The circle nodes correspond to subpathways of immune diseases. Two subpathways are connected by an edge if they share common genes. These subpathways are obtained by Subpathway-LDS method, using gene-phenotype associations of GAD. The topological properties including number of nodes and edges, average clustering coefficient, network diameter and average degree are shown right to the network. **(B)** Degree distribution of nodes. The X-axis and Y-axis refer to degree of nodes and numbers of nodes with certain degree respectively. **(C)**Clustering coefficient distribution of nodes. The Y-axis and X-axis refer to clustering coefficient and number of nodes with certain clustering coefficient respectively. **(D)** Distribution of shortest path length of the network. The X-axis denotes path length and the Y-axis denotes frequency of nodes in the network.

Figure 2. The global immune subpathway interaction network after adding subpathways of rheumatoid arthritis (GISI-RA network). The left side of the figure is the new immune network after adding subpathways of RA, in which the red nodes refer to the top ten risk subpathways of RA, the blue nodes refer to other subpathways of RA and the left gray nodes are subpathways of other immune diseases in the network. The right side is an amplification of the local region of network. Some important subpathways are marked with arrows and red texts.

Figure 3. Correlation scatter plot between random walk values and p-values of hypergeometric tests of the subpathways of rheumatoid arthritis identified by Subpathway-LDS method. The nodes denote subpathways of rheumatoid arthritis. The X-axis refer to the p-values of hypergeometric tests of the RA subpathways identified by Subpathway-LDS method through negative logarithmic transformation and Y-axis refer to 1000 times of rand walk scores.

Figure 4. Jak-STAT signaling pathway where the risk genes of rheumatoid arthritis were annotated. Nodes near asterisk symbol belong to the key subpathway region (RA-path:04630_1) identified by Subpathway-LDS. Enzymes (rectangular nodes) annotated by risk genes are showed with red node labels and borders.

Figure 5. Natural killer cell mediated cytotoxicity pathway where the risk genes of rheumatoid arthritis were annotated. Nodes near asterisk symbol belong to the key subpathway region (RA-path:04650_1) identified by Subpathway-LDS. Enzymes (rectangular nodes) mapped by risk genes are shown with red node labels and borders.

Supplementary Figure S1. MAPK signaling pathway where the risk genes of rheumatoid arthritis were annotated. Nodes near asterisk symbol belong to the key subpathway region (RA-path:04010_1) identified by iSubpathwayMiner. Enzymes (rectangular nodes) mapped by risk genes are shown with red node labels and borders.

Supplementary Figure S2. Prostate cancer pathway where the risk genes of rheumatoid arthritis were annotated. Nodes near asterisk symbol belong to the key subpathway region (RA-path:05215_1) identified by iSubpathwayMiner. Enzymes (rectangular nodes) mapped by risk genes are shown with red node labels and borders.

Supplementary Figure S3. Chemokine signaling pathway where the risk genes of rheumatoid

arthritis were annotated. Nodes near asterisk symbol belong to the key subpathway region (RA-path:04062_1) identified by iSubpathwayMiner. Enzymes (rectangular nodes) mapped by risk genes are shown with red node labels and borders.

Tables

Table 1. The top ten rheumatoid arthritis risk sub-pathways identified by Subpathway-LDS.

Subpathway Id	Pathway Name	Gene number annotated	Gene Number in pathways	P value
RA-path:04510_1	Focal adhesion	64	173	6.94E-09
RA-path:04010_1	MAPK signaling pathway	60	183	2.36E-06
RA-path:05414_2	Dilated cardiomyopathy	19	35	7.20E-06
RA-path:04540_1	Gap junction	27	62	9.50E-06
RA-path:04270_6	Vascular smooth muscle contraction	16	27	1.24E-05
RA-path:04512_9	ECM-receptor interaction	19	37	1.94E-05
RA-path:04012_1	ErbB signaling pathway	26	61	2.25E-05
RA-path:04020_1	Calcium signaling pathway	50	154	2.34E-05
RA-path:04512_6	ECM-receptor interaction	18	35	3.39E-05
RA-path:04512_10	ECM-receptor interaction	16	31	0.00010344

Table 2. The top ten rheumatoid arthritis risk sub-pathways optimized by random walk strategy based on immune disease sub-pathway crosstalk network.

Subpathway ID	Pathway name	Gene number	P value	Score
RA-path:04062_1	Chemokine signaling pathway	37	0.1053	0.0019
RA-path:04630_1	Jak-STAT signaling pathway	19	0.1791	0.0017
RA-path:05200_3	Pathways in cancer	51	0.0003	0.0017
RA-path:00830_1	Retinol metabolism	6	0.8188	0.0016
RA-path:04650_1	Natural killer cell mediated cytotoxicity	18	0.0564	0.0016
RA-path:04722_1	Neurotrophin signaling pathway	28	0.0006	0.0015
RA-path:00230_1	Purine metabolism	31	0.0707	0.0015
RA-path:04010_1	MAPK signaling pathway	60	2.36E-06	0.0015
RA-path:05215_1	Prostate cancer	21	0.0038	0.0014
RA-path:04020_1	Calcium signaling pathway	50	2.34E-05	0.0014

Table 3. The top ten first neighbors of MAPK signaling pathway ranked in ascending order of Fisher p values used for evaluating the significan of overlap genes shared by two different subpathways.

No.	Subpathway ID	P values of Fisher tests
1	lupus-erythematosus-path:04010_3	1.13E-44
2	arthritis-path:04010_1	4.84E-44
3	Crohn's-disease-path:04010_1	2.08E-43
4	periodontitis-path:05218_1	1.31E-34
5	arthritis-path:05218_2	9.12E-34
6	chronic-obstructive-pulmonary-disease-path:04010_2	2.50E-29
7	rheumatic-disease-path:04010_1	6.66E-28
8	sclerosis-path:04010_1	1.04E-27
9	Boeck's-sarcoid-path:04010_1	2.85E-26
10	diabetes-type-I-path:04010_2	1.20E-25

Table 4. The top ten first neighbors of Jak-STAT signaling pathway ranked in ascending order of Fisher p values used for evaluating the significance of overlap genes shared by two different subpathways.

No.	Subpathway ID	P values of Fisher tests
1	diabetes-type-I-path:04630_1	1.39E-127
2	Graves'-disease-path:04630_1	1.39E-127
3	allergic-diseases-path:04630_1	2.61E-120
4	asthma-path:04630_1	2.61E-120
5	atopic-dermatitis-path:04630_1	2.61E-120
6	atopy-path:04630_1	2.61E-120
7	Crohn's-disease-path:04630_1	2.61E-120
8	immune-globulin-path:04630_1	2.61E-120
9	arthritis-path:04630_1	1.53E-118
10	lupus-erythematosus-path:04630_1	1.53E-118

Table 5. The top ten first neighbors of chemokine signaling pathway ranked in ascending order of Fisher p values used for evaluating the significance of overlap genes shared by two different subpathways.

No.	Subpathway ID	P values of Fisher tests
1	Crohn's-disease-path:04062_1	4.08E-185
2	diabetes-type-I-path:04062_1	4.08E-185
3	asthma-path:04062_1	4.36E-157
4	arthritis-path:04062_1	2.74E-152
5	rheumatic-disease-path:04062_1	2.74E-152
6	organ-transplant-path:04062_1	5.64E-143
7	chronic-obstructive-pulmonary-disease-path:04062_1	2.74E-139
8	pancreatitis-path:04062_1	2.74E-139
9	periodontitis-path:04062_1	8.49E-134
10	atopic-dermatitis-path:04062_1	3.66E-121

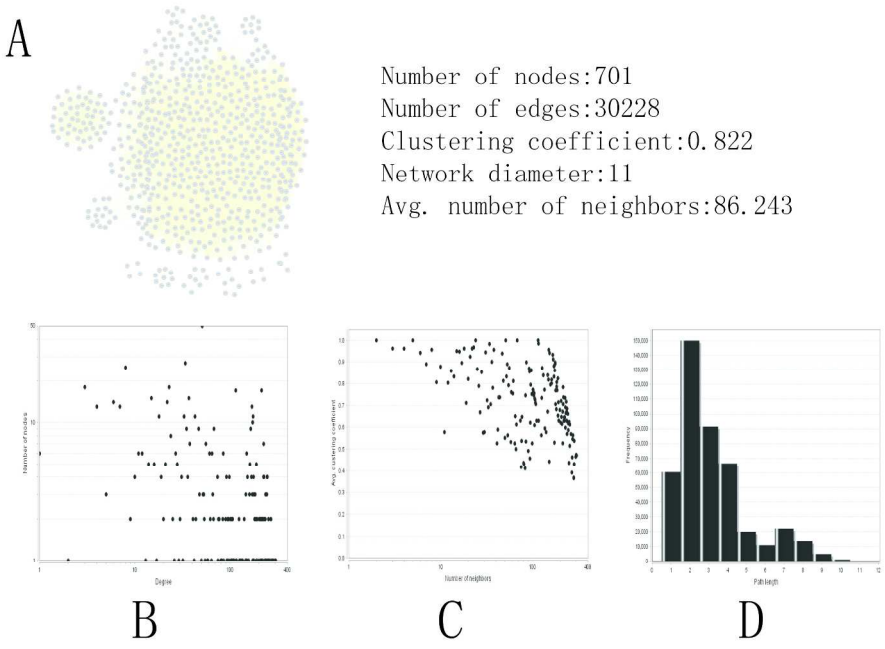


Figure 1. The global immune disease subpathway crosstalk network and fundamental topological properties. (A) The subpathway crosstalk network of 85 immune diseases. The circle nodes correspond to subpathways of immune diseases. Two subpathways are connected by an edge if they share common genes. These subpathways are obtained by Subpathway-LDS method, using gene-phenotype associations of GAD. The topological properties including number of nodes and edges, average clustering coefficient, network diameter and average degree are shown right to the network. (B) Degree distribution of nodes. The X-axis and Y-axis refer to degree of nodes and numbers of nodes with certain degree respectively. (C)Clustering coefficient distribution of nodes. The Y-axis and X-axis refer to clustering coefficient and number of nodes with certain clustering coefficient respectively. (D) Distribution of shortest path length of the network. The X-axis denotes path length and the Y-axis denotes frequency of nodes in the network.
236x158mm (300 x 300 DPI)

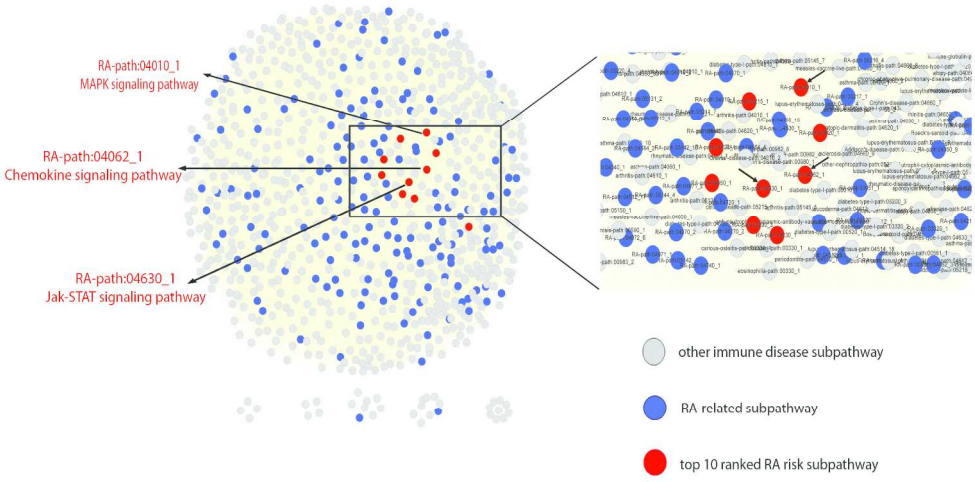


Figure 2. The new immune network after adding subpathways of rheumatoid arthritis. The left side of the figure is the new immune network after adding subpathways of RA, in which the red nodes refer to the top ten risk subpathways of RA, the blue nodes refer to other subpathways of RA and the left gray nodes are subpathways of other immune diseases in the network. The right side is an amplification of the local region of network. Some important subpathways are marked with arrows and red texts.
255x171mm (300 x 300 DPI)

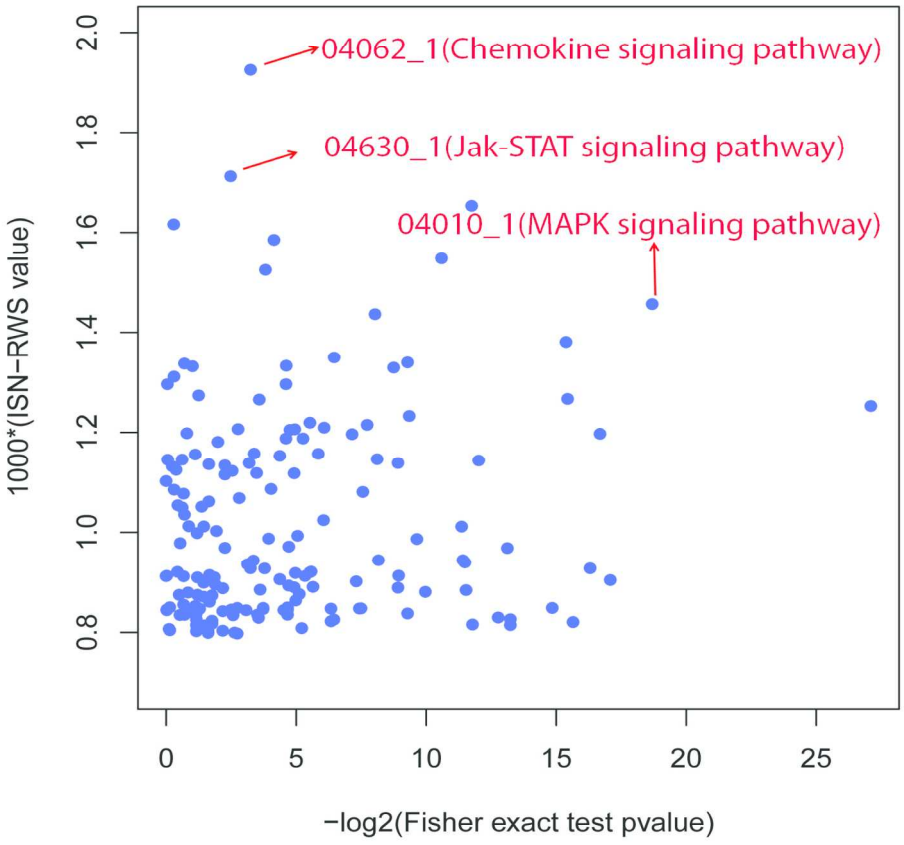


Figure 3. Correlation scatter plot between random walk values and p-values of Fisher exact test. The nodes denote subpathways of rheumatoid arthritis. The X-axis refer to the p-value through negative logarithmic transformation and Y-axis refer to 1000 times of rand walk scores.
181x167mm (300 x 300 DPI)

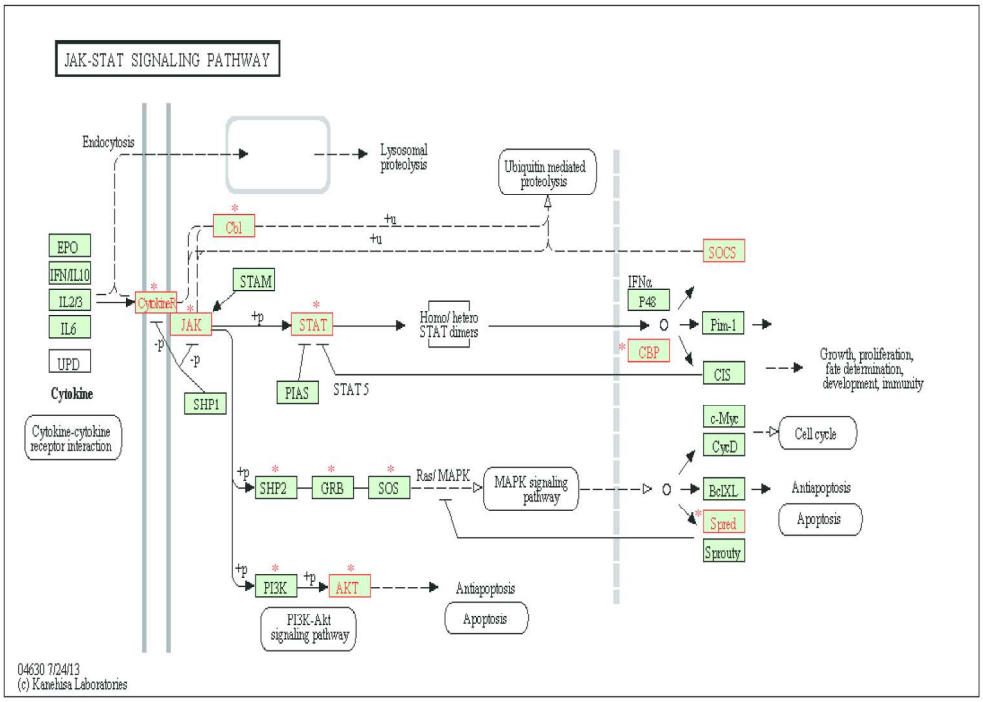


Figure 4. Jak-STAT signaling pathway where the risk genes of rheumatoid arthritis were annotated. Nodes near asterisk symbol belong to the key subpathway region (RA-path:04630_1) identified by Subpathway-LDS. Enzymes (rectangular nodes) annotated by risk genes are showed with red node labels and borders.
215x157mm (300 x 300 DPI)

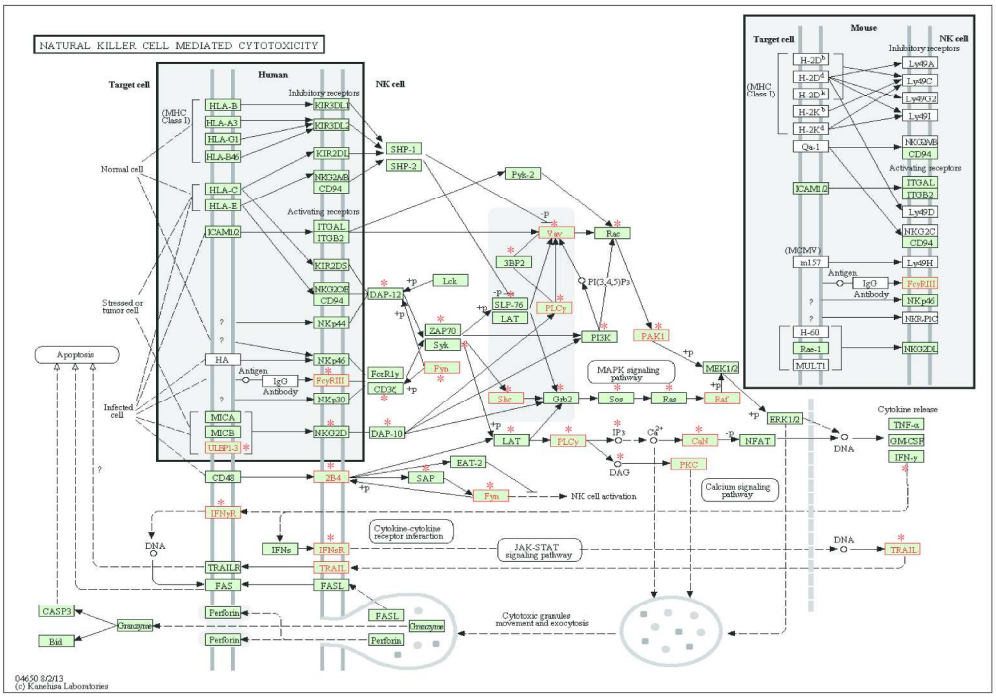


Figure 5. Natural killer cell mediated cytotoxicity pathway where the risk genes of rheumatoid arthritis were annotated. Nodes near asterisk symbol belong to the key subpathway region (RA-path:04650_1) identified by Subpathway-LDS. Enzymes (rectangular nodes) mapped by risk genes are shown with red node labels and borders.
289x205mm (300 x 300 DPI)