1. Introduction

Genome-wide association studies (GWAS) have successfully detected a larger number of genetic variants associated with the susceptibility of complex diseases (Saxena et al., 2007; T.W.T.C.C. Consortium, 2007). In traditional GWAS, thousands of genetic variants such as single nucleotide polymorphisms (SNPs) are tested simultaneously by comparing the allelic frequencies of each genetic variant between the normal and affected individuals. However, traditional approaches to identify genetic variants have limitations. First, very few genetic variants exceed the stringent significance threshold imposed by adjusting for the number of statistical tests performed. As a result, many variants that confer modest disease risk may be ignored. Second, a single variant that is closely associated with a complex disease may have only a small role in disease causation. Complex diseases often arise from the joint effect of multiple genetic variants (Peng et al., 2009). With the increasing number of GWAS coming online and the fact that single genes account for only a part of disease risk, understanding the joint effect of modestly associated genetic variants from GWAS is increasingly important. Several studies proposed that the combination of polymorphisms in several components of a biological pathway might significantly influence complex diseases (Low et al., 2010; Ulrich et al., 2003). Thus, examining the joint effect of genes that are modestly associated with disease has become an essential focus in complex disease research.

Recently, pathway-based approaches to analyze GWAS data have been proposed to study the joint effect of modestly disease-related genetic variants (Baranzini et al., 2009; Torkamani et al., 2008; Wang et al., 2007). Many disease-related pathways have been identified. However, most of the methods only consider the probability of disease-related genes co-occurring in pathways and do not consider the complex relationships between genes in pathways. Many studies showed that pathway structure information, such as the interactions and distances between gene products, can improve understanding of delicate pathway functions (Antonov et al., 2008; Draghici et al., 2007; Koyuturk et al., 2004; Ogata et al., 2000). Therefore, we can use the pathway structure information which considers the complex relationships between gene products to better analyze joint effect of genetic variants in pathways. The widely used pathway database, Kyoto Encyclopedia of Genes and Genomes (KEGG) which provides a reference knowledge base for linking genomes to life through the process of pathway mapping, can infer higher-level functions of disease processes and organism behaviors from high-throughput experimental technologies (Kanehisa et al., 2008). Moreover, KEGG database which consists of a collection of manually drawn pathway

Abstract

Most methods for genome-wide association studies (GWAS) focus on discovering a single genetic variant, but the pathogenesis of complex diseases is thought to arise from the joint effect of multiple genetic variants. Information about pathway structure, such as the interactions and distances between gene products within pathways, can help us learn more about the functions and joint effect of genes associated with disease risk. We developed a novel sub-pathway based approach to study the joint effect of multiple genetic variants that are modestly associated with disease. The approach prioritized sub-pathways based on the significance values of single nucleotide polymorphisms (SNPs) and the interactions and distances between gene products within pathways. We applied the method to seven complex diseases. The result showed that our method can efficiently identify statistically significant sub-pathways associated with the pathogenesis of complex diseases. The approach identified sub-pathways that may inform the interpretation of GWAS data.

© 2012 Elsevier B.V. All rights reserved.

Methods paper

Identifying disease related sub-pathways for analysis of genome-wide association studies

Chunquan Li 1, Junwei Han 1, Desi Shang 1, Jing Li, Yan Wang, YingYing Wang, Yunpeng Zhang, Qianlan Yao, Chunlong Zhang, Kongning Li *, Xia Li *

College of Bioinformatics Science and Technology and Bio-pharmaceutical Key Laboratory of Heilongjiang Province, Harbin Medical University, Harbin 150081, China

Keywords: GWAS Pathogenesis of complex diseases Pathway structure information Joint effect

ARTICLE INFO

Article history:
Accepted 18 April 2012
Available online 30 April 2012

Abbreviations: GWAS, genome-wide association studies; SNPs, single-nucleotide polymorphisms; RA, rheumatoid arthritis; CD, Crohn’s disease; T1D, type 1 diabetes; BD, bipolar disorder; HT, hypertension; CAD, coronary artery disease; T2D, type 2 diabetes; KEGG, Kyoto Encyclopedia of Genes and Genomes; KOs, KEGG Orthology identifiers; WTCCC, Wellcome Trust Case Control Consortium.

* Corresponding authors at: College of Bioinformatics Science and Technology, Harbin Medical University, 194 Xuefu Road, Harbin 150081, China. Tel.: +86 451 86615922; fax: +86 451 86615922.

E-mail addresses: kongningli@hotmail.com (K. Li), lixia@hrbmu.edu.cn (X. Li).

1 Contributed equally to this work.

© 2012 Elsevier B.V. All rights reserved.
SNPs were mapped to the corresponding sub-pathways through the Wellcome Trust Case Control Consortium (WTCCC). For each disease, the numbers of SNPs with a p-value less than 0.05 were considered significantly associated with the disease. The results are listed in Supplementary Table 2. To examine in detail the magnitude of difference scores for real and randomized p-values of seed nodes, we calculated for each disease the average score of all sub-pathways (Fig. 2(a)) and the score of the top 20 sub-pathways (Figs. 2(b)−(h)). Notably, the average scores were substantial higher from real data than from randomized data, suggesting that the high-scoring sub-pathways found by our method may be relevant to disease.  

2.2. Significant sub-pathways for RA

In the GWAS data for RA, 25,315 SNPs with p-values less than 0.05 were extracted. The SNPs mapped to 3910 genes. Using our method, we found 1331 seed nodes and 16 sub-pathways (Table 1) that were statistically significant (p < 0.05). In 16 sub-pathways, there are nine sub-pathways whose entire pathways they include in were also found to be statistically significant (p < 0.05) by an entire pathway identification method (hypergeometric test). Many studies suggest that these nine pathways such as purine metabolism (path: 00230), tight junction (path: 04530), and calcium signaling (path: 04020) are relevant to the pathogenesis of RA (Crupi et al., 2010; Davies and Hallett, 1998; Ewert et al., 2010; Forrest et al., 2006). For example, calcium (Ca²⁺) signaling was found to play an important role in controlling a diverse range of cellular processes, such as gene transcription, muscle contraction, and cell proliferation (Bootsman et al., 2001). The Ca²⁺ concentration of peripheral blood T lymphocytes was found to be significantly reduced in RA patients (Carruthers et al., 1996). Therefore, the calcium signaling pathway (path: 04020) may be significantly associated with the pathogenesis of RA. Our approach identified not only a sub-pathway (path: 04020.7) in calcium signaling but also its seed nodes. The seed nodes close to Ca²⁺ in the calcium signaling pathway may play a major role in Ca²⁺ signal transduction and in signaling calcium release (Fig. 3).  

More importantly, we found that 7 of the 16 sub-pathways that were statistically significant by our method occurred in pathways that were not statistically significant based on an entire pathway identification method (hypergeometric test) (p > 0.05). The seven pathways were alanine, aspartate, and glutamate metabolism (path: 00251); glycine, serine, and threonine metabolism (path: 00260); Jak-STAT signaling pathway (path: 04630); one carbon pool by folate (path: 00670); pyrimidine metabolism (path: 00240); valine, leucine and isoleucine biosynthesis (path: 00290); and regulation of actin cytoskeleton (path: 04810). The results for the seven sub-pathways are described below.

The alanine, aspartate and glutamate metabolism (path: 00251) contained one statistically significant sub-pathway. The sub-pathway had five seed nodes, namely K01580, K01915, K00823, K00764 and K00820 (Table 2). These gene products contained SNPs that may contribute modestly to disease but failed to reach the stringent threshold (p = 1.09 × 10⁻⁷) imposed by the correction for multiple testing. In the sub-pathway, the longest distance between these gene products was no more than three. This indicated that the gene products (enzymes) catalyze collective metabolic reactions. By mapping the gene products back to the original pathway, we found that the products are involved in glutamate metabolism (Fig. 4). The gene products were glutamate decarboxylase (GAD), glutamate-ammonia ligase (GLUL), 4-aminobutyrate aminotransferase (ABAT), phosphoribosylpyrophosphate amidotransferase (PPAT), and glucosamine-fructose-6-phosphate aminotransferase (GPT) (Table 2). In glutamate metabolism, GAD converts glutamate to 4-aminobutanoate (GABA). ABAT then converts GABA to succinic semialdehyde (Windmueller, 1982; Wu, 1998). The GABAergic system includes GAD and GABA, is located in many peripheral non-neural tissues, and may participate in controlling the peripheral non-neural tissues, and may participate in controlling the
inflammatory response in RA (Kelley et al., 2008; Tamura et al., 2009). GLUL converts glutamate to glutamine. The abnormal variant of GLUL may change the metabolism progression from glutamate to glutamine. Glutamine is an ingredient in a nutritional treatment for rheumatoid cachexia (Marcora et al., 2005). In addition, glutamate released by osteoblasts modulates bone tissue resorption, which plays an intriguing role in initiation and progression of RA (Hajati et al., 2009). These observations suggest that the collective actions of enzymes (GAD, GLUL and ABAT, etc.) may contribute to the initiation and progression of RA.

In the pathway for glycine, serine, and threonine metabolism (path: 00260), the sub-pathway mediating a major part of glycine and serine metabolism was statistically significant. Some seed nodes, such as serine hydroxymethyltransferase (SHMT) and cystathionine beta-synthase (CBS), catalyze sequential reactions. SHMT reversibly converts glycine to serine. CBS converts serine to cystathionine and L-cysteine (Chen et al., 2004; Florio et al., 2011). Cystathionine increases greatly in older RA patients (Partsch et al., 1977), and CBS gene transfection is associated with inflammatory response (Sen et al., 2007). SHMT affects the therapeutic efficacy and toxicity of methotrexate in RA patients (Andrade, 2009). Furthermore, serine and glycine were related to some special proteins involved in the immune/inflammatory response of RA (Baboonian et al., 1989; Hofmann et al., 2002; Toussirot and Roudier, 2007). These
observations suggest that these seed nodes (SHMT and CBS) may jointly play an important role in the initiation and progression of RA.

Our method also found statistically significant sub-pathways in the Jak-STAT signaling pathway (path: 04630); the one carbon pool by folate (path: 00670); pyrimidine metabolism (path: 00240); valine, leucine, and isoleucine biosynthesis (path: 00290); and regulation of actin cytoskeleton (path: 04810) (Table 1). These pathways have been linked to RA by others. For example, Walker and colleagues suggested modulation of the Jak-STAT pathway as a therapeutic strategy for RA (Walker and Smith, 2005). Furthermore, the seed nodes such as IL2RB (K05136), L22RA (K05069), JAK2 (K04447), and STAT4 (K11222) are good therapeutic targets for RA (Fridman et al., 2010; Kurreeman et al., 2009; Palomino-Morales et al., 2010). In addition, folate plays a key role in normal cell growth and replication (Beaudin and Stover, 2007) and the folate pathway inhibitors, methotrexate and 5-fluorouracil, have been used in various treatment regimens for RA (Ranganathan et al., 2008; Sokka et al., 2008). To sum up, our method identified statistically significant sub-pathways and seed nodes that may jointly play an important role in the pathogenesis of RA.

2.3. Relationships for seven complex diseases

Results for six other complex diseases: CD, T1D, BD, HT, CAD, and T2D are listed in Supplementary Table 2. We identified 45 sub-pathways and 565 seed nodes that may be associated with at least one disease. The significance values ($p$-values) of seed nodes associated with different diseases varied widely. To assess the similarity of the results among the diseases, we performed hierarchical clustering of the diseases based on Pearson correlation coefficients of the $p$-values of seed nodes (Fig. 5). T1T and CD clustered almost perfectly with each other and grouped with RA. These three diseases are primarily autoimmune disorders. BD and HT formed a second group; these diseases have strong neurological and metabolic components.

3. Discussion

The information available from the Human Genome Project has become increasingly comprehensive, so a better interpretation of GWAS data has become especially important for understanding the
The traditional methods for analyzing GWAS data normally focus on single genetic variants and have identified many SNPs that play important roles in disease pathogenesis. However, complex diseases are thought to arise from the joint effect of multiple factors. Most of the traditional GWAS methods neglect SNPs that are only modestly associated (p < 0.05) with disease and do not consider the joint effect of multiple genetic variants. Variants within the same pathway may jointly create a higher likelihood of diseases (Collins et al., 2003; Thomas, 2005). Our sub-pathway based approach focuses on the joint effect of SNPs that are modestly associated (p < 0.05) with complex diseases. We developed to rank SNPs and the distances between seed nodes in the sub-pathway.

We found the local areas of pathways after mapping the seed nodes of sub-pathways to the original pathways. The seed nodes may jointly contribute to disease-related biological functions. For example, the pathway for alanine, aspartate, and glutamate metabolism (path: 00251) was not statistically significant with RA in an entire pathway identification. However, the sub-pathway representing glutamate metabolism was statistically significant. In the sub-pathway, glutamate is respectively catalyzed by glutamate decarboxylase (GAD) and glutamate-ammonia ligase (GLUL) to 4-aminobutanoate (GABA) and glutamine. GABA is then catalyzed by ABAT to succinic semialdehyde. The abnormal forms of these enzymes which jointly change the progression of glutamate metabolism may participate in controlling the inflammatory response in RA.

To identify the sub-pathways that may be significantly associated with complex diseases, we used a novel strategy to identify the sub-pathways that may be significantly associated with complex diseases. The traditional approaches to GWAS data analysis can be applied to other published GWAS results to classify complex diseases and improve understanding of the pathogenesis of complex diseases.
4. Material and methods

4.1. Reconstructing pathways to graph models

The KEGG pathway database consists of a collection of manually drawn pathway maps (Fig. 6(a)) (Kanehisa and Goto, 2000; Kanehisa et al., 2008). In KEGG, all pathways which include: metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases and drug development pathways can be represented as a graph (Fig. 6(b)) (Koyuturk et al., 2004), and the pathway structure information can be represented in the graph. In order to construct pathway graphs, we extracted the

<table>
<thead>
<tr>
<th>Seed nodes</th>
<th>Enzymes</th>
<th>Gene symbols</th>
<th>SNPs ID (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>K00764</td>
<td>Phosphoribosyl pyrophosphate amidotransferase</td>
<td>PPAT</td>
<td>rs4865080 (5.16E−07)</td>
</tr>
<tr>
<td>K01915</td>
<td>Glutamate-ammonia ligase</td>
<td>GLUL</td>
<td>rs12563980 (0.04)</td>
</tr>
<tr>
<td>K00820</td>
<td>Glucosamine-fructose-6-phosphate aminotransferase</td>
<td>GPPT1</td>
<td>rs412863 (1.09E−05)</td>
</tr>
<tr>
<td>K01580</td>
<td>Glutamate decarboxylase</td>
<td>GAD1</td>
<td>rs357199 (0.017), rs17091340 (0.024), rs1041639 (0.025), rs641226 (0.04), rs1350648 (0.07), rs546741 (0.005), rs2114213 (0.004), rs6568863 (0.009), rs1954600 (0.015), rs1537184 (0.005), rs1486113 (0.003), rs17092178 (0.03), rs3128557 (0.004), rs2630389 (0.0036).</td>
</tr>
<tr>
<td>K00823</td>
<td>4-Aminobutyrate aminotransferase</td>
<td>GAD2</td>
<td>rs11015025 (0.002), rs7068510 (0.04).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ABAT</td>
<td>rs1641001 (0.008), rs1273351 (0.009).</td>
</tr>
</tbody>
</table>

Table 2
The seed nodes of the sub-pathway in alanine, aspartate and glutamate metabolism (path: 00251). The enzymes, genes and SNPs which correspond to these seed nodes are also shown.

Fig. 4. The alanine, aspartate and glutamate metabolism (path: 00251) where the seed nodes of rheumatoid arthritis (RA) were annotated. The general gene products (KOs) were colored purple. The seed nodes in the submitted gene products (KOs) were colored red. The results show that these seed nodes were mostly concentrated in local areas of the pathway which corresponds to major part of glutamate metabolism.
relationships among gene products which are represented by KOs from the XML files of pathways in KEGG (ftp://ftp.genome.jp/pub/kegg/xml/ko). As metabolic pathways differ from other pathways in structure, we adopted different strategies to extract the relationships among gene products (KOs) for the two types of pathways. Undirected graphs were constructed in the following ways: (i) for metabolic pathways, two KOs were connected by an edge if there was a common compound in the KOs’ corresponding reactions (Li et al., 2009); (ii) for other pathways in KEGG database, two KOs were connected by an edge if they had relationships such as binding and modification indicated in the relation element of the XML file. Through the above operations, all pathways were converted to undirected graphs with KOs as nodes, and reactions or relationships as edges (Fig. 6(b)). Since many sub-graphs were not connected, we extracted all the connected sub-graphs for further use. The sub-graphs containing more than two nodes were considered as sub-pathways (Fig. 6(c)). If two KOs are connected by a limited number of steps in a pathway, they will exist in the same sub-pathway. The distance between two arbitrary KOs was calculated as the shortest path from one KO to the other in a sub-pathway.

4.2. Identifying seed nodes

The WTCCC performed GWAS for seven complex diseases (RA, CD, T1D, BD, HT, CAD, and T2D) (T.W.T.C.C. Consortium, 2007). For each disease, thousands of SNPs are tested simultaneously by comparing the allelic frequencies of each SNP between the normal and affected

---

Fig. 5. Hierarchical cluster of diseases by p-values of seed nodes (KOs) in disease related sub-pathways. p-Values of KOs associated with each disease are varied widely across diseases. Diseases were then clustered on these p-values of seed nodes by using distance measures based on the Pearson correlation. This procedure was performed by cluster and the graphic was prepared using TreeView [Eisen et al., 1998].

---

Fig. 6. The process of the pathway is converted to sub-pathways. (a) A pathway in KEGG. (b) The pathway is converted to the undirected graph with gene products as nodes. (c) The connected sub-graphs which contain more than 2 nodes are defined as the sub-pathways.
individuals. The p-values of association with the disease for each tested SNP are calculated. However, very few SNPs exceed the genome-wide significance threshold because of the adjustment for multiple comparisons and many SNPs that confer modest disease risk may be ignored. The joint effect of multiple modestly disease-related genetic variants will play a significant role in the development of disease (Peng et al., 2009). To study the interplay of modestly disease-related SNPs, we downloaded the p-values of association for each tested SNP from the WTCCC website (http://www.wtccc.org.uk/). All SNPs with these p-values less than 0.05 were selected and mapped to the corresponding genes. The lowest p-value of all SNPs located in a given gene was assigned as the gene-wise significance value; this approach was shown to be effective in handling the problem with genes having different number of SNPs (Baranzini et al., 2009; Torkamani et al., 2008). These genes which include at least one SNP with p < 0.05 can be used to identify significant entire pathways. In this work, the statistical significance of entire pathway identification method was evaluated by a hypergeometric test. Similarly, we assigned the lowest p-value of all genes encoding a given KO as the significance value of the KO. The KOs were defined as seed nodes. The seed nodes with at least one disease-related SNP may be associated with disease. We converted the KO-wise p-value to a z-score using \( z_i = \phi^{-1}(1 - p_i) \), where \( \phi^{-1} \) is the inverse normal CDF (cumulative density function). Thus, p-values were distributed uniformly from 0 to 1, z-scores followed a standard normal distribution, and lower p-values corresponded to larger z-scores. The seed nodes and their z-scores were used as major factors in our metric ranking the association of sub-pathways with disease.

4.3. Identifying high-risk sub-pathways

In pathways, the functional similarity between two KOs tends to increase as the distance between the KOs decreases (Guo et al., 2006; Ogata et al., 2000). Therefore, two KOs that are close to each other may perform similar biological functions. The KOs in tightly connected sub-pathways may likewise have highly similar functions. The distance between each pair of KOs can be expressed as the shortest path between them in the corresponding sub-pathway. If one sub-pathway contains many significant seed nodes that are close to each other, these seed nodes may jointly play a role in the pathogenesis of a complex disease. Considering the number of seed nodes, the p-values of seed nodes, and the distance between each pair of seed nodes, we developed a novel measure to rank the disease-risk of a sub-pathway as follows:

\[
S = \max_{i \neq j} \left\{ \sqrt{k} \left( \sum_{j \neq i} \frac{1}{k} \sum_{j \neq i} z_j \right) \right\}
\]

In this formula, S is the score of one sub-pathway, v is the set of seed nodes in the sub-pathway, k is the number of seed nodes, and \( d_{ij} \) is the shortest path between a pair of seed nodes (i and j). A high S indicates that the sub-pathway may be associated with the pathogenesis of complex diseases.

To prove that high sub-pathway scores are reliable indicators of disease association, we performed a permutation analysis to quantify the distribution of scores that may arise by chance. As the metric, S is based on the z-scores and distances of seed nodes in each sub-pathway, we shuffled seed nodes across all nodes in the set of sub-pathways. With this approach, the number of seed nodes remained constant but the location of seed nodes in the sub-pathways occurred by chance. To generate the background distribution, k permutations were performed. In each permutation, sub-pathway scores were calculated. To estimate the significance of a given sub-pathway, the real score, S, of the sub-pathway was compared to the background distribution. Let m represent the number of values equal to or greater than S in the background distribution. The p-value of the sub-pathway was computed as p-value = m/k. A low p-value indicates a higher likelihood of association between a sub-pathway and a given disease.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 30871394, 30370798, 30900837, and 30571034), the National High Tech Development Project of China, the 863 Program (Grant No. 2007AA022329), the National Basic Research Program of China, the 973 Program (Grant No. 2008CB517302), and the National Science Foundation of Heilongjiang Province (Grant Nos. ZJG501, 1055HG009, GB03CG02-4, and BMFH060044).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2012.04.051.

References