A systematic method for mapping multiple loci: An application to construct a genetic network for rheumatoid arthritis

Chuanxing Li, Guangmei Zhang, Xia Li, Shaoqi Rao, Binsheng Gong, Wei Jiang, Dapeng Hao, Peng Wu, Chao Wu, Lei Du, Yun Xiao, Yingying Wang

Department of Bioinformatics, the First Clinical College, and the Bio-pharmaceutical Key Laboratory of Heilongjiang Province and State, Harbin Medical University, Harbin 150081, PR China

Department of Bioinformatics, Capital University of Medical Sciences, Beijing 100084, PR China

Department of Computer Science, Harbin Institute of Technology, Harbin 150080, PR China

Department of Medical Statistics and Epidemiology, School of Public Health, Sun Yat-Sen University, Guangzhou 510080, PR China

Received 2 April 2007; received in revised form 1 October 2007; accepted 19 October 2007

Available online 4 November 2007

Abstract

The advent of high-throughput single nucleotide polymorphisms (SNPs) omics technologies has brought tremendous genetic data. Systematic evaluation of the genome-wide SNPs is expected to provide breakthroughs in the understanding of complex diseases. In this study, we developed a new systematic method for mapping multiple loci and applied the proposed method to construct a genetic network for rheumatoid arthritis (RA) via analysis of 746 multiplex families genotyped with more than five thousands of genome-wide SNPs. We successfully identified 41 significant SNPs relevant to RA, 25 associated genes and a number of important SNP–SNP interactions (SNP patterns). Many findings (loci, genes and interactions) have experimental support from previous studies while novel findings may define unknown genetic pathways for this complex disease. Finally, we constructed a genetic network by integrating the results from this analysis with the rapidly accumulated knowledge in biomedical domains, which gave us a more detailed insight onto the RA etiology. The results suggest that the proposed systematic method is powerful when applied to genome-wide association studies. Integrating the analysis of high-throughput SNP data with knowledge-based SNP functional annotation offers a promising way to reverse engineer the underlying genetic networks for complex human diseases.

© 2007 Elsevier B.V. All rights reserved.

Keywords: Genome-wide association; Integrative strategy; Ensemble decision; Interaction; Complex disease

Abbreviations: ADAM10, a disintegrin and metalloprotease domain 10; ADAMTS17, a disintegrin-like and metalloprotease (reprolysin type) with thrombospondin type 1 motif, 17; ALDH9A1, aldehyde dehydrogenase 9 family, member A1; ALS2, amyotrophic lateral sclerosis 2 (juvenile); BTBD7, BTB (POZ) domain containing 7; CD160, CD160 molecule; CFTR, cystic fibrosis transmembrane conductance regulator; FAM40B, family with sequence similarity 40, member B; FAM63B, family with sequence similarity 63, member B; FN, false negative; FP, false positive; FSHR, follicle stimulating hormone receptor; GAW15, Genetic Analysis Workshop 15; IBD, identical (identity) by descent; IL-1, The protein Interleukin 1; LOC493754, RAB guanine nucleotide exchange factor (GEF) 1 pseudogene; MLF1IP, MLF1 interacting protein; NARAC, North American Rheumatoid Arthritis Consortium; OMIM, On-line Mendelian Inheritance in Man; PCMT1, protein-L-isoaspartate (D-aspartate) O-methyltransferase; PDZK1, PDZ domain containing 1; PLAU, plasminogen activator, urokinase; RA, rheumatoid arthritis; RUNX1, Runt-related transcription factor 1; SLC9A4, Solute carrier family 9 (sodium/hydrogen exchanger), member 4; SNP, single nucleotide polymorphism; SNTG1, syntrophin, gamma 1; TNF, Tumor necrosis factors, a class of cytokines TP, true positive; UTR, untranslated regions; WDR62, WD repeat domain 62; ZNF140, zinc finger protein 140.

* Corresponding authors. Department of Bioinformatics, Harbin Medical University, Harbin 150080, PR China. Tel.: +86 451 8661 5922; fax: +86 451 8661 5922.

E-mail addresses: lixia6@yahoo.com (X. Li), shaoqir@yahoo.com (S. Rao).

0378-1119/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
1. Introduction

Identifying inherited genetic variation and genes involved in a disease is crucial to fully understand its pathology. In the past few years, large-scale SNP studies have promised some success in identifying genes and genomic regions involved in development of various disease phenotypes. However, it is not obvious how to analyze the omics data most efficiently. Most traditional analytical techniques consider each genetic marker separately, and thus are less efficient in providing a deep understanding of the sophisticated interplays between these genetic risk factors. A lesson learnt from the increasing evidence coming from model organisms and human studies, suggests that interactions among multiple loci contribute broadly to complex traits. Systematic evaluation of high dimensional SNPs is thus expected to provide some breakthroughs towards understanding the complexities involved in diseases. In this study, we developed a new systematic method for mapping multiple disease loci. We then applied the method to construct a genetic network for rheumatoid arthritis (RA) via analysis of 746 multiplex families genotyped with more than five thousands of genome-wide SNPs. We successfully identified 41 significant SNPs relevant to RA, 25 associated genes and a number of important SNP–SNP interactions (SNP patterns). Many findings (loci, genes and interactions) have experimental support from previous studies while novel findings may define unknown genetic pathways for this complex disease. Finally, we constructed a genetic network by integrating the results from this analysis with the rapidly accumulated knowledge in biomedical domains, which gave us a more detailed insight onto the RA etiology.

2. Materials and methods

2.1. SNP data

The SNP data analyzed in this study were provided by the Genetic Analysis Workshop 15 (GAW15; http://www.gaw15.org), which were initially investigated by the North American Rheumatoid Arthritis Consortium (NARAC) and consisted of 746 multiplex Caucasian RA families. Prior to analysis, we selected the familial individuals and markers according to the following rules. First, each individual had a clear-cut diagnosis for RA, either affected or unaffected. Second, at least one SNP marker was genotyped for the individual. Third, all the selected SNPs were from autosomes. Therefore, the final data used for this analysis contained 1989 individuals including 1640 affected and 349 unaffected individuals, and 5407 SNP markers. Because of the low missing rate (no individual with > 8% missing genotypes across the SNP markers and no SNP with > 4% missing genotypes across all the individuals), we did not impute the missing values.

2.2. Data pre-processing

We used a two-stage design to conduct the genome-wide association analysis. First, we reduced the dimensionality of SNPs by following an entropy-based and an IBD (Identity By Descent)-based data pre-processing procedure. Entropy, proposed by Shannon, measures the uncertainty of random variables and is a useful measure for DNA variation analysis. Zhao et al. (2005) proved that the entropy difference of a SNP between affected and unaffected individuals can reflect the associated strength between the marker and the studied disease. Here, the entropy was defined as:

$$S(X) = E[-\log P(x)] = -\sum_i P(x_i)\log P(x_i),$$

where \(X\) denoted a SNP genotype or status of RA, \(P(x_i)\) was the probability that \(X\) had the value of \(x_i\). The mutual information between the marker \((X)\) and the studied disease \((Y)\), which measured the amount of information shared by the two variables, was defined as:

$$I(X, Y) = S(X) + S(Y) - S(X, Y),$$

where \(S(X)\) and \(S(Y)\) were the entropies of variables \(X\) and \(Y\), respectively, and \(S(X, Y)\) was their joint entropy.

We implemented a three-step, entropy-based data pre-processing procedure. First, SNPs were grouped into four genotypic categories: heterozygote, two homozygotes and an unknown genotype. Next, based on the concepts of entropy and mutual information, we designed four indices to evaluate the effect of a particular marker on the disease: (1) the mutual information between marker genotypes and disease status; (2) the difference between two entropy values of a marker in affected and unaffected populations, (3) the difference between two entropy values of a marker in all samples and the affected populations; and (4) the difference between two entropy values of a marker in all samples and the normal population. In order to estimate the cut-off value at a significant level \((P=0.05)\), we then constructed the respective empirical distributions of the four indices by randomly sampling 100,000 times from the values of all markers using a bootstrap method, a similar permutation method to the described in our previous work (Li et al., 2004), which was under the premise that most of the genome-wide SNPs were unrelated to a disease phenotype. This analysis identified a total of 550 distinct SNPs (247, 310, 272 and 281 for four indices, respectively), where the cut-off values at \(P=0.05\) were 0.0062, 0.0014, 0.0663 and 0.0785, respectively for the four indices.

Our previous work (Li et al., 2005) demonstrated that the SNP IBD difference between concordantly affected and concordantly unaffected sib pairs could be used to identify the association between a SNP and a disease. We computed the SNP IBD data using the GENIBD program of the S.A.G.E. package (S.A.G.E., 2005). The IBD difference of a single marker was depicted by the discrepancy in its two mean IBD values for groups of concordantly affected and concordantly unaffected sib pairs, respectively. We also randomly sampled 100,000 times from the values of all markers to construct a corresponding empirical distribution, and this analysis identified 289 significant SNPs at \(P=0.05\) with the cut-off value of 0.1863. Finally, by fusing the results obtained by using the above strategies, we got a marker set of 787 unique SNPs for...
further data analysis. In this study, to avoid the possible loss of
the true positives, we did not perform a multiple-test correction
for the number of SNPs evaluated. Instead, we employed a step-
wise procedure to control the overall Type I error rate during
implementing the proposed multi-stage approach. Therefore,
the \( P \)-value quoted in each step should be considered as a
heuristic measure, useful as an indicator for roughly rating
relative importance of the candidate SNPs.

2.3. Identifying powerful decision trees

In complex diseases, a combination or interaction of multiple
SNPs, plus environmental factors, may better determine the
associated disease susceptibility. Here we utilized decision trees
to evaluate the differentiating ability of individual SNPs as well
as their combinations (i.e., SNP patterns) (Li et al., 2004; Xie
et al., 2005) for separating disease from other phenotypes. A
classification tree was constructed using the Matlab-decision
tree toolbox, where each SNP was a node variable. We used an
“IF–THEN” rule (Xie et al., 2005) and Gini’s diversity index as
the criteria for choosing a split. The tree growth was stopped if a
further split at the current node did not improve the purity of its
child nodes or when there were less than 10 observations in an
impure node.

We generated 2000 classification models (decision trees) using
a bootstrapping method. Each model was trained and tested with a
fixed sample size of training (1380 samples) and test set (340
samples) and with the same proportion of affected and unaffected
individuals in both training and test sets (59% versus 41%). We
used two indices to evaluate the power of the tree: chi-square and
\( F \)-value (\( F \)-value = \( 2 \times TP / (2 \times TP + FP + FN) \)), where TP, FP and
FN represent the number of true positives, false positives and false
negatives, respectively. We then selected 320 powerful trees for
further identification of relevant SNPs and SNP patterns, each
tree with \( \chi^2 \)-square and \( F \)-value \( \geq 3.84 \) (degree of freedom = 1; \( P \leq 0.05 \) and \( F \)-value \( \geq 0.6509 \) \( P \leq 0.05 \)), which was determined by randomly
sampling 100,000 times from the \( F \)-values of all trees to construct
the corresponding empirical distribution.

2.4. Ensemble decision for identifying important SNPs and
their interactions

Here, we extended our previously proposed ensemble method
(Li et al., 2004), which was demonstrated via analysis of the gene
expression profiles of colon cancer providing more reliable and
robust molecular signatures for predicting complex disorders. The
basic rationales followed two rules. First, a SNP may be more
important to a disease if it was chosen by more trees and if it was in
the upper layer of a tree (closer to the root). As shown in Eq. (1),
we designed a weighted index \( SW_i \) for SNP \( i \) that was the sum of
its contributions in \( n \) trees where \( SW_i^k \) described its contribution in the
\( k \)-th tree and was related to its depth \( d_i^k \) in the tree (the depth of the
root node was zero). Second, each branch or path on a tree
implicated a decision rule. Thus, the frequency of two or more
SNPs (SNP pattern) occurring in the same paths can measure their
interaction intensity (joint contribution). To this end, we defined a
weighted index \( (PW) \) to describe SNP interactions in a pair-wise
manner. As shown in Eq. (2), this index measured the concordance
rate of a SNP pair among all the paths in \( n \) trees, where \( m_k \) was the
number of branches in the \( k \)-th tree.

\[
SW_i = \sum_{k=1}^{n} SW_i^k = \sum_{k=1}^{n} 2^{-d_i^k + 1}
\]

\[
PW_{ij} = \sum_{k=1}^{n} m_k \sum_{l=1}^{m_k} p_{wij}^l,
\]

\[
p_{wij}^l = \begin{cases} 1, & \text{if both SNPs } i \text{ and } j \text{ appear in path } l \text{ of the } k \text{th tree.} \\ 0, & \text{otherwise.} \end{cases}
\]

Next, we randomly sampled 100,000 times from the \( SW \) values of all markers using a bootstrap method to construct the
corresponding empirical distribution, and then defined the cut-off
value (\( SW = 213 \)) at a significant level \( (P = 0.05) \) for mining
disease relevant SNPs. SNP patterns were ranked by \( PW \) values
and the top 100 SNP patterns were selected for further analysis in
the network construction.

3. Results

Using our proposed method, we analyzed the SNP data of RA
and obtained 787 markers from data pre-processing, 320 power
trees and identified 41 significant SNPs. The top 100 SNP patterns
were selected for further analysis in the network construction.

3.1. Identification of 41 SNPs relevant to RA

We applied our method to the SNP data for RA and identified
41 significant markers. By integrating the information from the
GAW15 investigation and several public databases including
dbSNP (Sherry et al., 1999), OMIM (Schorderet, 1991),
Pubmed (Richardson, 1998) and Mapview (http://www.ncbi.
nlm.nih.gov/mapview/), we derived the associated genes, the
corresponding loci (cytobands) and some supportive evidence
on their involvements in RA (see Table 1 for detail). Then we
evaluated this SNP set regarding its biological implications. The
identified SNP set showed a high relevance to RA biology. As
shown in Table 1, 16 of the 41 SNPs had been found associated
with RA based on the previous studies. The results strongly
suggested that these SNPs may confer either high risks or
protective roles to RA.

3.2. Identification of 25 associated genes relevant to RA

By searching in dbSNP, we found 21 of the 41 SNPs to be
closely linked to 25 known genes (Table 1), of which SNPs
were within the intron regions and seven SNPs were at the
untranslated regions (UTR) of genes. Although the function of
some SNPs (e.g. within an intron or UTR) waits to be
characterized, these SNPs may point to the intron and UTR
regions that are regulatory in terms of modulating transcripti-
onal activities of down-stream genes or modifying the
structures of gene products. Among the 25 genes, four genes were found to be associated with RA, including RUNX1, CFTR, PLAU and ADAM10. Among the remaining 21 genes (potential genetic factors for RA), many have either direct or indirect relationships (e.g. regulatory or binding) with the above four known RA genes. There were many intense cross-talks between the 25 identified genes, including SNP–SNP interactions, regulatory and/or binding relationships.

3.3. Construction of a RA-relevant network

Among the 320 powerful decision trees, 116,883 different interacting SNP pairs were found. We evaluated the RA-driven interactions by using Eq. (2). The reversely-engineered RA-relevancy network (partially shown in Fig. 1) consisted of two types of nodes and two types of edges. Both dash- and solid-line nodes denoted the genes identified by using our method, which were further dug out using the significant SNPs as landmarkers (Table 1). A solid-line edge described a SNP–SNP interaction and the circled number was its rank among all 116,883 pair-wise patterns. Here we only showed SNP patterns with ranks less than 100. A dash-line edge represented a relationship between two genes in the nodes, both of which could interact with the gene or gene product in the elliptical circle. These dash-line edges were presented here mainly for demonstrating the potential SNP–SNP or gene–gene interactions learnt from the SNP–SNP patterns. Further analysis of this network revealed two interesting findings. First, in terms of connectivity (the number of links with a gene), the top two nodes in this network were ALS2 and CFTR. Gene ALS2 (amyotrophic lateral sclerosis 2 (juvenile)) is a validated protein-coding gene, which has a role in regulating Rac-PAK signalling and neurite outgrowth and is involved in juvenile amyotrophic lateral sclerosis (MIM: 205100). Seen from the network in Fig. 1, it was also one of the most important RA-susceptibility genes and had strong interactions with both RUNX1 and CFTR.

Table 1

<table>
<thead>
<tr>
<th>Ranka</th>
<th>SNP IDb</th>
<th>Genec</th>
<th>Locusd</th>
<th>OMIM and referencese</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs970595</td>
<td>ALS2</td>
<td>2q33.1</td>
<td>(Osorio et al., 2004; Plenge et al., 2005; Amos et al., 2006), CTLA4</td>
</tr>
<tr>
<td>2</td>
<td>rs1004531</td>
<td></td>
<td>5q23.1</td>
<td>(Plenge et al., 2005)</td>
</tr>
<tr>
<td>3</td>
<td>rs164466</td>
<td></td>
<td>3p25.3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>rs1001250</td>
<td></td>
<td>2q33.1</td>
<td>(Osorio et al., 2004; Plenge et al., 2005; Amos et al., 2006), CTLA4</td>
</tr>
</tbody>
</table>

aThe rank of SNP according to index SW in Eq. (1).
bSNP identifier in dbSNP database.
cSNP-associated gene derived from dbSNP database.
dCytoband localized by the linked SNP.
eAdditional support evidence from the OMIM database (entries shown), literature or the related genes (e.g. CTLA4 and SLAM).

Among the 320 powerful decision trees, 116,883 different interacting SNP pairs were found. We evaluated the RA-driven interactions by using Eq. (2). The reversely-engineered RA-relevancy network (partially shown in Fig. 1) consisted of two types of nodes and two types of edges. Both dash- and solid-line nodes denoted the genes identified by using our method, which were further dug out using the significant SNPs as landmarkers (Table 1). A solid-line edge described a SNP–SNP interaction and the circled number was its rank among all 116,883 pair-wise patterns. Here we only showed SNP patterns with ranks less than 100. A dash-line edge represented a relationship between two genes in the nodes, both of which could interact with the gene or gene product in the elliptical circle. These dash-line edges were presented here mainly for demonstrating the potential SNP–SNP or gene–gene interactions learnt from the SNP–SNP patterns. Further analysis of this network revealed two interesting findings. First, in terms of connectivity (the number of links with a gene), the top two nodes in this network were ALS2 and CFTR. Gene ALS2 (amyotrophic lateral sclerosis 2 (juvenile)) is a validated protein-coding gene, which has a role in regulating Rac-PAK signalling and neurite outgrowth and is involved in juvenile amyotrophic lateral sclerosis (MIM: 205100). Seen from the network in Fig. 1, it was also one of the most important RA-susceptibility genes and had strong interactions with both RUNX1 and CFTR. For gene CFTR (cystic fibrosis transmembrane conductance regulator), Doyen et al. found its association with RA (Doyen et al., 2005).

![Fig. 1. A reversely-engineered RA-relevant network.](image-url)
Second; the frequency measure PW did efficiently capture the joint contributions of two SNPs toward RA. For example, based on the PW measures of the linked SNPs, ALS2 was strongly associated with CFTR (ranked the 31st). Molecular evidence shows that both ALS2 and CFTR are able to bind to RAB5A (Hyink et al., 2001; Otomo et al., 2003). It is also known that CFTR regulates NFKB and NFKB that in turn regulate PLAU (Wolf et al., 1993; Weber et al., 2001), rendering a regulatory path of ALS2-CFTR-PLAU, which also partially explained a lower rank (the 2454th) for the SNP–SNP pair corresponding to genes ALS2 and PLAU, possibly owing to their indirect interaction separated by CFTR. Overall, the constructed RA-relevant network (see Fig. 1) revealed several most important known pathogenic hub genes and interactions, and also some unknown hubs and pathways or part of several connected pathways (e.g. from ALS2 to RUNXI) waiting for further investigations.

3.4. Methodological validation

To validate our proposed method, we compared its performance with 2-by-n contingency table chi-square test, which is commonly used to assess the population-based association between markers and a disease, and Haseman–Elston (H–E) regression-based linkage test, which can be used to identify the non-random association due to the genetic linkage between two genomic loci. For the H–E linkage test, the provided pedigree data of 746 families were analyzed using the SIBPAL program of the S.A.G.E. package (S.A.G.E., 2005). For the chi-square test, an independent dataset (constructed by sampling one individual per family) was analyzed.

For point-wise analysis to identify significant markers, a 2-by-3 trait-genotype table was analyzed by the chi-square test (or Fisher Exact Test if necessary), as implemented in the R package (http://www.r-project.org/). Of the three SNP sets of 5407 SNPs (denoted as “All SNPs” in Table 2), 787 candidate SNPs selected by the data pre-processing and 41 identified SNPs, we found significant (P ≤ 0.05) associations at 297 (5.49%), 123 (23.63%) and 31 (75.61%) markers, respectively. These results indicated that our proposed method was largely consistent with the traditional chi-square test in identifying significant association with RA. However, nearly one quarter (10) of SNPs were identified only by our method, but not by chi-square test. Further analysis showed that three of the 10 markers were linked to some known RA related genes (e.g. CFTR, PLAU and RUNXI). This fact may largely reflect the power of a tree model in identifying some important interacting features although their marginal contribution might not be salient as revealed by chi-square test. To our surprise, the significance results from the H–E linkage test were overlapped with other two methods in small proportions (e.g. at P = 0.05, between the H–E test and chi-square test: the overlapping SNPs were 33, 14, 4, respectively for the three SNP sets; between the H–E test and our proposed method: the overlapping rates were 102, 5, respectively for the SNP sets of 787 candidate SNPs and 41 identified SNPs). Although several variations, for example, the genetic principles based, the data utilized, and inherent methodological differences, may explain the lack of overlapping in significance, further comprehensive investigation is warranted.

To explore the pair-wise SNP patterns using an alternative method, we constructed a 2-by-9 genotype table for each SNP pair. In some cases when a cell is scare, genotype collapsing was performed. We tested the SNP pairs in three SNP sets: (1) 787 candidate SNPs, with 309,291 pairs constructed; (2) the SNPs included in 320 decision trees, with a total of 116,883 SNP pairs both appearing at least in one tree path; (3) the set of 41 SNPs identified by using SW, which gives 783 SNP pairs appearing at least in one tree path. For comparison, the top 100 SNP pairs for (2) and (3) were also defined by PW. For the three SNP sets tested, the chi-square test found 46,455 (15%), 28,135 (24%) and 620 (79.18%) SNP pairs of significance (P ≤ 0.05), respectively (Table 3), indicating that our proposed approach is effective in finding majority (up to about 80%) of SNP–SNP interaction as defined by 2-by-n contingency table. Among the top 100 SNP pairs for (2) and (3) defined by PW, at the significance level of 0.05 there were 26 and 92 SNP pairs cross-validated by the alternative method. Compared with the previous point-wise analysis, the percentage of cross-validated SNP pairs in the SNP set of 41 SNPs identified by using SW was slightly increased from 75.61% to 79.18%. This promising result suggests that the indirect method for identifying SNP–SNP interaction following the prior point-wise assessment of single SNPs using SW is a feasible alternative for mining multiple loci, without much risk of losing some markers significantly interacting with others as SW is defined under the environment of complex interaction pathways.

4. Discussion and conclusion

Single nucleotide polymorphisms (SNPs) are the most amenable type of polymorphic marker to a large-scale and high-density analysis on a genome-wide level. The analysis of SNPs permits determining relationships between genotypes and

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Numbers of SNPs of significant association with RA revealed by 2-by-3 contingency table chi-square test or H–E regression-based linkage test</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP set</td>
<td>Chi-square test</td>
</tr>
<tr>
<td></td>
<td>P ≤ 0.05</td>
</tr>
<tr>
<td>All SNPs (5407)</td>
<td>297</td>
</tr>
<tr>
<td>Candidate SNPs (787)</td>
<td>123</td>
</tr>
<tr>
<td>Identified SNPs (41)</td>
<td>31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Numbers of SNP pairs of significant association with RA, revealed by 2-by-n contingency table chi-square test</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP pair set</td>
<td>No. pairs</td>
</tr>
<tr>
<td>Between candidate SNPs</td>
<td>309,291</td>
</tr>
<tr>
<td>SNP pairs identified by PW</td>
<td>116,883</td>
</tr>
<tr>
<td>Top 100 SNP pairs</td>
<td>100</td>
</tr>
<tr>
<td>Between SNPs identified by SW</td>
<td>783</td>
</tr>
<tr>
<td>Top 100 SNP pairs</td>
<td>100</td>
</tr>
</tbody>
</table>
phenotypes as well as identification of SNPs related to a disease. Identifying genes involved in disease progression is crucial to fully understand its molecular pathology, and for developing novel therapeutics. The advent of high-throughput SNP technologies has brought an on-slaught of genetics profiles aiming at identifying some inherited genetic variations that have critical but as yet largely uncharacterized roles in development of human disease (The International HapMap Consortium, 2005). Searching for disease relevant SNPs as the landmark(s) to locate disease gene(s) is a critical step for positional cloning of the underlying molecular determinants for complex human traits. In this study, we presented a new and systematic bioinformatic genome-wide analysis strategy for mapping multiple loci followed by construction of a disease-driven genetic network. This strategy exploited the respective strengths of a traditional genetic epidemiological approach (i.e. valid inference based on familial resemblance) and a bioinformatics approach (i.e. capability of finding useful interaction patterns in high dimension data). Application to the RA data demonstrated its efficiency and success in mining genome-wide SNP data. Many of the findings (SNP associated loci, genes and interactions) from this study have already gained convincing support from various previous studies in biomedical domains. We also constructed a RA-relevant network, based on the interaction intensity measuring the joint contribution of paired SNPs toward differentiating the health status of an individual, which provided us with a broader and deeper view on the RA etiological topology.

Despite large-scale SNP studies in recent years, especially the most recent reports for whole genome scans using newly developed large-scale industrial genotyping platforms (e.g. (Herbert et al., 2006; Easton et al., 2007; Helgadottir et al., 2007; McPherson et al., 2007; Wellcome Trust Case Control Consortium, 2007)), have demonstrated some success in identifying genes and genomic regions involved in development of various disease phenotypes, it still remains an open issue to analyze the omics data most efficiently and to meet a list of challenges (Heidema et al., 2006) when systematic evaluation of thousands and millions of SNPs at the omics level is performed. Based on the analytical strategies, current methods for genome-wide SNP association studies can be categorized into two groups. First group of methods use a one-step approach incorporated with a well-developed genetic epidemiological based algorithm, often making asymptotic distribution-based statistical estimation and inference and controlling overall Type I error rate using Bonferroni-type correction of multiple tests on massive SNPs. Second group of methods are referred to the recently developed integrated multiple-stage based methods that employ various statistical and data mining approaches (Longmate, 2001; Ionita and Man, 2006), aiming at providing solid and robust estimation and inference under best control of the overall (experimental-wide, genome-wide or analysis-wide) type I error and meanwhile without loss of power for detecting the real underpins based on the limited sample sizes collected to human genetic association studies.

Our newly proposed method is multiple-stage based consisting of the following steps: (1) a set of candidate SNP markers are identified using the entropy and the IBD measure, largely for screening out less informative or noise features; (2) a number of decision trees are constructed based the set of candidate markers and the bootstrap to identify powerful combinations of features aimed to assess their individual contributions and joint actions with others; (3) a final set of candidate markers are identified based on their roles and importance among the decision trees; (4) the genetic network is constructed based on the pair-wise patterns and joint contributions of the SNPs appearing in the decision trees and the information for the “position-altered” genes and the gene–gene interactions via knowledge mining of the rapidly accumulated databases. During development of the integrative method and its application to the SNP data, we have devoted consistent efforts and attentions to build a statistically sounding support for the proposed analysis strategy. Because of the distribution-free nature of the data mining approaches integrated and after reviewing current advances in tackling multiple-tests issues in analyzing large-scale SNP data, we proposed the multiple-stage design and a step-wise control of the overall false positives. For testing the association between a single SNP and the studied disease under the null hypothesis that no association between the allelic variations and the occurrence of the disease(s), we used a nominal significance of 0.05 in steps 1–3, which generated an overall type I-error of the range from $0.05^{\sim}=1.25 \times 10^{-3}$ to $0.05^{\sim}=6.25 \times 10^{-6}$ (if the two tests (chi-square and F-value test) in step 2 were considered independent), slightly liberal than the corresponding Bonferroni-corrected rate of $9.25 \times 10^{-6}$ (0.05/5407) for a one-step approach. However, in step 4 and for testing gene–gene interactions that are deemed to be harder to test, we used a more strict criterion of $\sim 10^{-3}$ (100/100,000 pairs). Based on this study and our previous study for analysis of simulated datasets for alcoholism (Li et al., 2006), the proposed multiple-steps approach with a step-wise control of Type I error is proved effective in identifying either most of simulated true loci or disease relevant biomarkers.

It should be pointed out that most of data mining algorithms often cannot explicitly take into account the potential issue of non-independence between observations when applied to pedigree data. In this study, we ignored this issue when computing and analyzing entropy values based on our empirical analysis of its impact on analysis of the studied pedigrees consisting of mostly small size families, and because of the lack of suitable data mining approaches to dealing with the pedigree-induced residual correlation structure. To investigate the issue, we constructed an ‘independent’ dataset (essentially by sampling one individual per family) from the original pedigree data to assess the influence of pedigree structure on the entropy computation. Using the entropy values for both the original data and the derived independent data, we calculated the Pearson’s correlation coefficient ($r$): (i) between the entropy values for the original pedigree data and for the independent data ($r=0.9955$); (ii) between the entropy values for the original normal-phenotype data and the independent normal-phenotype data ($r=0.9951$); and (iii) between the entropy values for the original disease data and the independent disease data ($r=0.9914$).
These results show that the entropy values for the original pedigree data and for the independent data resemble remarkably each other, indicating whether or not take into the issue of non-independence had no significant impact on the entropy computation for this particular dataset. The particular data structure characterized by small family sizes may explain the fact of essentially no discrepancies between the entropy values computed when the issue of non-independence for the data was taken or not taken into account. About 82.8% families (629 families) are of 2 to 3 members with genotyping data, and 11 (1.47%) families are of one genotyped member.

However, though ignoring the non-independence issue was feasible for this particular analysis, it should be noticed that ignoring such dependence among samples would generally result in inflated (or reduced) type I error rates, and so we should be cautious about the impact of non-independence when general pedigree data are analyzed by using the proposed integrated approach. Otherwise, some remedy measures to remove the potential confounding due to the residual familial covariance should be taken prior to imputing the SNP genetic features into the proposed machine learning toolbox.

Finally, we should recognize the limited power of any single association study to uncover the comprehensive list of important biomarkers and/or genes for a complex disease. First, the likelihood of the lack of a major gene(s) and the involvement of many modest-effect genes acting in concert can result in inflated (or reduced) type I error rates, and so we should be cautious about the impact of non-independence when general pedigree data are analyzed by using the proposed integrated approach. Otherwise, some remedy measures to remove the potential confounding due to the residual familial covariance should be taken prior to imputing the SNP genetic features into the proposed machine learning toolbox.

Finally, we should recognize the limited power of any single association study to uncover the comprehensive list of important biomarkers and/or genes for a complex disease. First, the likelihood of the lack of a major gene(s) and the involvement of many modest-effect genes acting in concert can lead their effects undetectable based on the sample sizes often recruited in human genetics studies. On the basis of reviewing nearly 500 papers on disease association studies and a number of genome-wide linkage scans, Ober and Hoffjan (2006) estimated that the total number of genes that contribute to risk of asthma may exceed 100 and that the individual effect of any one of these genes on disease risk is quite small. Second, although there are few cases to support the so-called “common disease/common variant hypothesis” (e.g. common variants in the apolipoprotein E (APOE, MIM # 10774) in Alzheimer and diabetes population) (Dean, 2003), the frequencies for the (functional) genetic variants identified in candidate-gene studies are often too low and evaluation of the relationship between such rare events and the studied disease is error-prone. Third, many complex diseases (e.g. RA (Weyand et al., 1995)) and even previously deemed simple Mendelian characters are remarkably heterogeneous and the heterogeneities can be demonstrated at the levels of clinical phenotyping, genetics and environmental factors (Dean, 2003; Biesecker, 2004) and the presence of heterogeneities essentially violates the underlying assumptions for most of epistemiological approaches utilized in human genetic studies. Fourth, current SNP genotyping platforms may not provide adequate and genome-wide uniformed SNP covering with a sufficient density, leading to some important genes unfound in a single study. For instance, some well-known cytokine genes (IL-1, TNF, etc.) for RA mechanisms were not seen on the hit-table (Table 1). To investigate the reason, we looked up the NCBI dbSNP database and found 71 SNPs linked to IL-1 and 91 SNPs linked to TNF. However, none of the above SNPs were included in the 5407 SNPs analyzed. These facts may explain the lack of a finding for the two well-known genes. Nevertheless, it is interesting to note that we did identify gene ADAM10 (a disintegrin and metalloprotease domain 10) in our list, a member of ADAM protein family, which can cleave many proteins including TNF-alpha and E-cadherin.

In conclusion, this study demonstrated that an integrated analysis of high-throughput SNP data using the proposed bioinformatics toolbox and with knowledge-based SNP functional annotation offers a promising and robust way to uncover the molecular underpinnings and to reverse engineer the underlying genetic networks for complex human diseases. However, some limitations of the proposed data mining based multiple-steps method, especially when applied to pedigreedata, should be well recognized.

Acknowledgements

We thank the Genetic Analysis Workshop 15 organizers and the original investigators of the North American Rheumatoid Arthritis Consortium (NARAC) for providing the SNP data for this methodological study. This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 30600367, 30370798, 30571034 and 30570424), The National High Tech Development Project of China, the 863 Program (Grant Nos. 2007AA02Z329), National Science Foundation of Heilongjiang Province (Grant Nos. ZJG0501, 1055HG009, GB03C602-4, and BMFH060044). Some of the results of this paper were obtained by using the program package S.A.G.E., which is supported by a U.S. Public Health Service Resource Grant (RR03655) from the National Center for Research Resources.

References


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


