BMC Medical Genomics

RESEARCH ARTICLE Open Access

A towards-multidimensional screening approach to predict candidate genes of rheumatoid arthritis based on SNP, structural and functional annotations

Liangcai Zhang, Wan Li, Leilei Song, Lina Chen*

Abstract

Background: According to the Genetic Analysis Workshops (GAW), hundreds of thousands of SNPs have been tested for association with rheumatoid arthritis. Traditional genome-wide association studies (GWAS) have been developed to identify susceptibility genes using a "most significant SNPs/genes" model. However, many minor- or modest-risk genes are likely to be missed after adjustment of multiple testing. This screening process uses a strict selection of statistical thresholds that aim to identify susceptibility genes based only on statistical model, without considering multi-dimensional biological similarities in sequence arrangement, crystal structure, or functional categories/biological pathways between candidate and known disease genes.

Methods: Multidimensional screening approaches combined with traditional statistical genetics methods can consider multiple biological backgrounds of genetic mutation, structural, and functional annotations. Here we introduce a newly developed multidimensional screening approach for rheumatoid arthritis candidate genes that considers all SNPs with nominal evidence of Bayesian association ($BFLn > 0$), and structural and functional similarities of corresponding genes or proteins.

Results: Our multidimensional screening approach extracted all risk genes ($BFLn > 0$) by odd ratios of hypothesis H_1 to H_0, and determined whether a particular group of genes shared underlying biological similarities with known disease genes. Using this method, we found 6614 risk SNPs in our Bayesian screen result set. Finally, we identified 146 likely causal genes for rheumatoid arthritis, including CD4, FGFR1, and KDR, which have been reported as high risk factors by recent studies. We must denote that 790 (96.1%) of genes identified by GWAS could not easily be classified into related functional categories or biological processes associated with the disease, while our candidate genes shared underlying biological similarities (*e.g.* were in the same pathway or GO term) and contributed to disease etiology, but where common variations in each of these genes make modest contributions to disease risk. We also found 6141 risk SNPs that were too minor to be detected by conventional approaches, and associations between 58 candidate genes and rheumatoid arthritis were verified by literature retrieved from the NCBI PubMed module.

Conclusions: Our proposed approach to the analysis of GAW16 data for rheumatoid arthritis was based on an underlying biological similarities-based method applied to candidate and known disease genes. Application of our method could identify likely causal candidate disease genes of rheumatoid arthritis, and could yield biological insights that not detected when focusing only on genes that give the strongest evidence by multiple testing. We hope that our proposed method complements the "most significant SNPs/genes" model, and provides additional insights into the pathogenesis of rheumatoid arthritis and other diseases, when searching datasets for hundreds of genetic variances.

* Correspondence: chenlina_2004@yahoo.com.cn
Department of Biophysics, College of Bioinformatics Science and Technology; Harbin Medical University; Harbin, Hei Longjiang Province, China

Background

Rheumatoid arthritis is an inflammatory disease, primarily of the joints, with autoimmune features and a complex genetic component [1]. It arises from the underlying functional involvement of one or more mutated genes [1,2]. The essential challenge of rheumatoid arthritis is finding an effective screening approach to find candidate risk genes by their structural and functional similarity to known disease genes, and using them to develop new techniques for testing, diagnosis, and treatment [3-5].

When case-control datasets of complex diseases are available, genome-wide association studies (GWAS) have great power to detect genetic variants, especially if many markers are tested across the genome [6-8]. All published GWAS have led to the discovery of novel genes for complex diseases that differ between case and control groups. However, because of the arbitrary multiple testing used in these studies, genetic variants that confer a small disease risk but are of potential biological importance are likely to be missed using a "most significant SNPs/genes" approach [9,10]. To avoid the strict adjustment required in multiple testing, we developed a genome-wide Bayesian association method to test for association of a single SNP with a case-control phenotype. The Bayesian approach compares the probability of an association to the probability given no association. For complex diseases, discovering new bioinformatics strategies based on genome-wide Bayesian association methods that avoid the limitations of other study is vital.

Traditional statistical genetics aims to identify susceptibility genes based only on a statistical model without considering biological similarities between disease genes and likely causal genes. Proteins are essential parts of organisms and participate in virtually every cellular process. Most proteins fold into unique sequence arrangements and structures, and contribute to specific characteristics in diverse function sets. Proteins and genes that are responsible for complex diseases are often associated through similar sequences and structures [11-13], so candidate genes could be screened according to sequence, arrangement, and crystal structures that are similar to known disease genes. A support vector machine (SVM) is a machine learning algorithm based on Statistical Learning Theory that is commonly applied to resolve this problem [14-18]. Good classification effects can be obtained with only a few learning samples. Many studies [19-22] have demonstrated that disease genes with a specific phenotype share similar functionalities, and therefore, similarity in the functional annotations of these genes could be used to screen for candidate genes for a specific disease. A limited number of studies have used GWAS [23-25], function clustering algorithms [26-29], or machine learning methods based on structural genomics knowledge bases [30] to identify candidate genes for rheumatoid arthritis. When a set of candidate risk genes are acquired from case-control datasets of genetic variances, joint consideration of the structural and functional associations between candidate genes and a disease might provide additional insights into the results of traditional statistic genetics analysis for identifying candidate genes.

In this article, we hypothesized that underlying candidate genes harboring markers with minor or modest evidence of association could be identified through attributions they share with known disease genes, using multidimensional biological annotations such as gene sequence arrangement, crystal structure of encoded proteins, and similar biological pathways or mechanisms. Here, we introduce a newly developed multidimensional screening approach to predict candidate genes of rheumatoid arthritis based on SNPs, and structural and functional annotations. The rationale for performing our multidimensional candidate gene screen was the assumption that several genes, each modestly associated with a disease, may share sequence or structural pattern, and jointly participate in the same biological function to confer susceptibility. We used a genome-wide Bayesian association method to test for association between a case-control phenotype and a single SNP. To avoid the strict adjustment required for multiple testing, Bayesian approaches compare the probability of an association to the probability of no association. An SVM classifier was used to distinguish likely causal genes from non-disease genes by the sequence and crystal structural features of their proteins. Candidate genes were assumed to be disease genes if they were in the same functional categories or biological pathways associated with the pathogenesis. We carried out literature searches to verify our results, and compared them with traditional GWAS results to demonstrate the potential utility of this method.

Methods

Genetic Association Data of Rheumatoid Arthritis

Genotype frequencies of tested SNPs for case-control samples were downloaded from GAW16 online using the 500 K Affymetrix chip, from 868 cases and 1194 controls from the rheumatoid arthritis collection and normal samples http://www.gaworkshop.org/. Genotype frequencies were preprocessed to allele frequencies for each SNP.

Gene Location and Disease Loci Data

Location information for human genes was from the NCBI genome database (downloaded on Mar 25, 2009). Disease loci information was gathered from the OMIM online database (downloaded on Mar 25, 2009) [31].

Sequence and Crystal Structure Data

Linear-sequence items for all human genes were from the NCBI genome database. Crystal structure datasets of human proteomics were from online databases PDB http://www.rcsb.org/pdb/home/home.do and targetDB (http://targetdb.pdb.org/, downloaded on Mar 25, 2009).

Functional Annotations Data

Function categories in the PIRSF (http://pir.georgetown.edu/pirsf/, downloaded on Mar 31, 2009), GO (http://www.geneontology.org/, downloaded on Mar 31, 2009), and KEGG (http://www.genome.jp/kegg/, downloaded on Mar 31, 2009) databases were used as source function annotations, whose well-defined categories are widely used for important functional identification analysis. In this study, each candidate gene was annotated onto its corresponding functional families or categories using these three databases.

Genome-wide Bayesian Association Analysis

We assume here that data *D*, are counts of cases and controls for each of the three genotypes at a SNP locus (Table 1). Bayesian approaches compare the probability of *D* if there is an association (alternate hypothesis H_1) to its probability given no association (null hypothesis H_0). Although most case-control studies are retrospective, we adopted a prospective viewpoint in which a case-control status was the outcome variable and the genotype was regarded as known. Under H_0, the probability of the observed dataset D does not depend on genotype, and can be written in terms of the probability θ that an individual included in the study is a case,

$$P(D/\theta) = c\theta^{n^A}(1-\theta)^{n^U} \tag{1}$$

where we introduce n^A and n^U for the numbers of cases (affected) and controls (unaffected), and ^{C} is a combinatorial constant that cancels out below and so can be ignored. Here $^\theta$ is a "nuisance" parameter, whose value is not important, so under the Bayesian approach we eliminated it by integration with a prior probability distribution. For the purposes of this illustration, the uniform prior is a convenient choice, so

$$P(D) = \int_0^1 P(D/\theta)d\theta = cB(n^A+1, n^U+1) = B(n^A+1, n^U+1) \tag{2}$$

where B denotes the Beta function, defined by

$$B(n^A+1, n^U+1) = \frac{n^A!n^U!}{(n^A+n^U+1)!}$$

where $n^A! = n^A \times (n^A\text{-}1) \times (n^A\text{-}2) \times K \times 1$.

To compute a probability for D under H_1, we assumed that individuals with genotype *j* had a probability of θ_j to be a case. Then, analogous to (1),

$$P(D \mid \theta_0, \theta_1, \theta_2) = c{\theta_0}^{n_0^A}(1-\theta_0)^{n_0^U} \times \theta_1^{n_1^A}(1-\theta_1)^{n_1^U} \times \theta_2^{n_2^A}(1-\theta_2)^{n_2^U} \tag{3}$$

where ${n_j}^A$ and ${n_j}^U$ denote the numbers of cases and controls with genotype $j = 0, 1, 2$. We took the easiest approach first, assuming that each θ_j had an independent, uniform prior, and integrating to obtain [32]

$$\begin{aligned} P(D) = cB(n_0^A+1, n_0^U+1) \times B(n_1^A+1, n_1^U+1) \\ \times B(n_2^A+1, n_2^U+1) \end{aligned} \tag{4}$$

The next step was to compute the Bayes Factor (BF), which is the ratio of (4) to (2). The corresponding formula is:

$$BF = \frac{B({n_0}^A+1, {n_0}^U+1)B({n_1}^A+1, {n_1}^U+1)B({n_2}^A+1, {n_2}^U+1)}{B({n_0}^A+{n_1}^A+{n_2}^A+1, {n_0}^U+{n_1}^U+{n_2}^U+1)} \tag{5}$$

Values of BF larger than one support H_1, while BF < 1 indicated support for the null H_0.

To reduce the computational complexity, we used the log value of BF, *BFLn*, as our final function to screen the significant SNP set associated with the disease for each SNP V_i (i = *1,..., N*, where *N* is the total number of SNPs in the GWAS.

$$\begin{aligned} BFLn(V_i) = \ln B({n_0}^A+1, {n_0}^U+1) + \ln B({n_1}^A+1, {n_1}^U+1) \\ + \ln B({n_2}^A+1, {n_2}^U+1) - \ln B({n_0}^A+{n_1}^A \\ + {n_2}^A+1, {n_0}^U+{n_1}^U+{n_2}^U+1) \end{aligned} \tag{6}$$

Table 1 Frequency of cases and controls for each of three genotypes at a SNP locus.

Genotype:	AA	AB	BB	Total
Case	${n_0}^A$	${n_1}^A$	${n_2}^A$	${n_0}^A+{n_1}^A+{n_2}^A$
Control	${n_0}^U$	${n_1}^U$	${n_2}^U$	${n_0}^U+{n_1}^U+{n_2}^U$
Total:	${n_0}^A+{n_0}^U$	${n_1}^A+{n_1}^U$	${n_2}^A+{n_2}^U$	${n_0}^A+{n_1}^A+{n_2}^A+{n_0}^U+{n_1}^U+{n_2}^U$

Where ${n_0}^A, {n_1}^A, {n_2}^A, {n_0}^U, {n_1}^U, {n_2}^U$ represents the frequency for each of the specific genotype AA, AB or BB, respectively.

Values of $BFLn(V_i)$ larger than zero support H_1, while $BFLn(V_i) < 0$ indicated support for the null H_0.

We then associated gene g_t (t = *1... T*), where T was the number of all genes in the human genome, with SNP V_i, if this SNP was located within g_t or if g_t was the closest to V_i. SNPs that were 500 kb from any gene were considered because most enhancers and repressors are <500 kb away from genes, and most linkage disequilibrium blocks are <500 kb away [9]. We carried out first dimensional screening, namely genetic screening, by collecting a test set from genes associated with at least one significant SNP ($BFLn > 0$) and located within one or more disease loci for further filtering.

SVM Classification based on Sequence and Structure Similarity Features

ID Converter[33] was used to map all genes to their corresponding proteins across the entire human genome. In this section, the positive set consisted of rheumatoid arthritis disease genes from NCBI and the OMIM online database (Additional file 1). The negative set contained the remaining genes that did not fall within any disease loci after excluding genes in the positive set and the test set.

To simplify our analysis, a 28-dimension vector of physicochemical features (Table 2), a combinational pseudo-sequence, was used to represent each protein in positive, negative, and testing sets, according to the online RCSB PDB and targetDB databases. We used 8-dimension secondary features (21-28) and the entire 28-dimension physicochemical features to train two classifiersfor the second screen.

Considering the diversity of the putative non-disease-candidate proteins, the non-disease-candidate space might not have been sampled completely. Therefore, we constructed 1000 additional training sets (positive:negative = 1:1), in which each negative set was selected randomly from the original negative set. During each randomization, the 8- and 28-dimension features were used to construct the corresponding classifier. The performance of our model was evaluated with an n-fold cross-validation test. In the cross-validation test, the entire positive and negative data sets were shuffled and split into n folds. Each fold was used in turn for testing and the remaining part (n-1 folds) used for training. The sensitivity (Q_p), specificity (Q_n) and overall accuracy (Q_a) were used to measure the accuracy of positive prediction, negative prediction, and the overall accuracy of the model [34], respectively.

Table 2 Protein sequence and structure-based features from PDB and targetDB databases.

Dimension	Feature	Properties
1-20	C	Composition of the 20 amino acid residues
21	a	Cell length a in Angstroms
22	b	Cell length b in Angstroms
23	c	Cell length c in Angstroms
24	alpha	Cell angle alpha in degrees
25	beta	Cell angle beta in degrees
26	gamma	Cell angle gamma in degrees
27	helical	Percent of helical in protein sequence
28	beta sheet	Percent of beta sheet in protein sequence

$$Q_p = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$$
$$Q_n = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$$
$$Q_a = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})$$

Variables are true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). In general, the overall accuracy Q_a was used to measure the predictive power of a model.

We constructed 10,000 additional training sets (positive:negative = 335:335), in which each negative set was selected randomly from the original negative set. Here, the test set was prepared from genetic screens. We used the training sets, the test set, and the optima classifier to retrain the classifier, aiming to classify the genes in the test set, and then predict candidate genes for the disease. We performed randomization 10,000 times, and kept genes that were judged to be disease genes at each process of randomization.

The SVM used here was Libsvm http://www.csie.ntu.edu.tw/~cjlin/libsvm[35]. The commonly used kernel function, radial basis function (RBF) was introduced into our analysis. According to machine learning theory [36], an optimal hyperplane was drawn by the SVM model to separate positive samples from negative ones. The distance to the hyperplane is related to the confidence of a prediction. Therefore, the distance from each sample to the hyperplane was employed to predict the disease candidate likeness for genes or proteins.

Functional Annotation Screening and Candidate Gene Prediction

The PRISF, GO, and KEGG pathway databases are widely used for functional studies and gene annotations [37,38]. We hypothesized that disease genes would gather in specific protein families, participate in the same biological functions, or interact within specific biological pathways. According to these three databases, specific biological functions were annotated for known disease genes, and for genes from SVM screening. In this section, we define three functions (f_{PIRSF}, f_{GO} and f_{KEGG}) to evaluate whether each candidate gene in the set from SVM screening was strongly associated with the disease. The corresponding function formulas were:

$$f_{PIRSF}(g_i) = \begin{cases} 1, & \text{if } \mathrm{g}_i \text{ is annotated onto at least one} \\ & \quad \text{protein family that disease genes enriched;} \\ 0, & \text{otherwise;} \end{cases}$$

$$f_{KEGG}(g_i) = \begin{cases} 1, & \text{if } \mathrm{g}_i \text{ is annotated onto at least one} \\ & \quad \text{KEGG pathway that disease genes enriched;} \\ 0, & \text{otherwise;} \end{cases}$$

$$f_{GO}(g_i) = \begin{cases} 1, & \text{if } \mathrm{g}_i \text{ is annotated onto at least one} \\ & \quad \text{GO term that disease genes enriched;} \\ 0, & \text{otherwise;} \end{cases}$$

$$f(g_i) = f_{PIRSF}(g_i) \vee f_{GO}(g_i) \vee f_{KEGG}(g_i) = \begin{cases} 1, & if\ g_i \text{ is defined as a candidate;} \\ 0, & \text{otherwise.} \end{cases}$$

where g_i is any gene in the resulting set from the first and second screens.

Here, a sample description was listed below for functional annotation screening (Table 3).

Finally, we used function $f(g_i)$ for functional annotation screening by retaining genes ($f(g_i)$ = 1) that shared at least one similar functional annotation.

Comparison with Traditional GWAS

Traditional GWAS analysis [39] uses the Fisher exact test and multiple testing adjustment. Functional enrichments in GO biological processes and KEGG pathways were carried out for known disease genes, GWAS genes, and our predicted genes. Functional consistency with known disease genes was examined to evaluate GWAS genes and our predicted genes. To further evaluate the performance of our screening method, we used the NCBI PubMed module to retrieve associations of GWAS genes or our genes for rheumatoid arthritis using the term "GENE symbols+rheumatoid arthritis" (*e.g.* CD4+rheumatoid arthritis) to determine the underlying mechanisms of the genes from our model.

Results and Discussions

A genome-wide Bayesian association analysis was carried out to identify variants within genes that were modestly associated with rheumatoid arthritis. The dataset, produced by the Genetic Analysis Workshop (GAW), was 2062 samples genotyped with an Affymetrix Gene Chip Human Mapping 500 K Array Set. Quality control for this dataset included assessment of marker genotype frequency, allelic frequency, and departure from Hardy-Weinberg equilibrium. A total of 433,766 SNPs survived the quality control protocol and were tested for association with the trait in 868 cases and 1194 controls. Significant SNPs (*BFLn > 0*) were mapped onto their corresponding genes, and these genes were considered for further analysis. This process resulted in 4402 candidate risk genes, which were labeled as members of the test set for the SVM screening step.

Table 3 Sample description for functional annotation screening.

*Genes**	f_{PIRSF}	f_{GO}	f_{KEGG}	$f = f_{PIRSF}\ V_{GO}\ V_{KEGG}$
g_1	0	0	0	0
g_2	0	0	1	1
g_3	0	1	1	1
...	...	...	...	...
g_n	1	1	1	1

* contains the genes from SVM screening; columns f_{PIRSF}, f_{GO}, f_{KEGG}, are functional annotation information from PRISF, GO and KEGG databases; '*V*'is an "or" command.

For the SVM screening, we extracted the sequence and structure information from the PDB and targetDB databases and calculated their feature values for 335 known disease genes, 28,874 non-disease genes, and 4402 other genes in the test set. We used 8-dimension secondary features (21-28) and the entire 28-dimension physicochemical features to train two classifiers for the second screening (see Materials and Methods). To address the concern that, considering the diversity of the putative non-disease-candidate proteins, the non-disease-candidate space might not have been sampled completely, we constructed 1000 additional training sets (positive:negative = 1:1), with each negative set selected randomly from the original negative set. For each randomization, the 8- and 28-dimension features were used to construct the corresponding classifiers. The performance of our model was evaluated with a 5-fold cross-validation test in which the entire positive and negative data sets were shuffled and split into five folds. Each fold was used for testing, and the remaining part (5-1 folds) was used for training. The 1000 randomization results from the two classifiers were analyzed, and the relevant accuracy of 28-dimension physicochemical features varied between 0.695 and 0.891 (Table 4).

Based on predictions from these two classifiers, we chose the second classifier for candidate gene prediction. We reconstructed 10,000 additional training sets (positive:negative = 335:335), in which each negative set was selected randomly from the original negative set. The test set was prepared from genetic screens. After 10,000 randomizations, the intersection of each prediction was defined as the final prediction, resulting in 495 candidate genes that used for the third screening step.

We used three functional databases (PRISF, Gene Ontology [GO], and KEGG pathway) to identify

Table 4 Performance information of two classifiers based on 8-dimension secondary physicochemical features and 28-dimension physicochemical features.

features	prediction	average ± std
8-dimension secondary features(21-28)	0.631-0.831	0.703 ± 0.038
28-dimension physicochemical features	0.695-0.891	0.762 ± 0.036

responsible risk genes that had similar functional annotations to known disease genes. We hypothesized that disease genes would gather in specific protein families, participate in the same biological functions, and interact in specific biological pathways. According to the three databases, specific biological functions were annotated for 335 known disease genes and 495 genes from the SVM screening. For 495 genes, according to defined function $f(g_i)$, we collected candidate genes for which each value of their corresponding function $f(g_i)$ equaled 1, demonstrating their strongly functional associations with known disease genes. We identified 146 candidate disease genes as our final candidate predictions for rheumatoid arthritis (Additional file 2).

We used the web software toolkit, Gene Webgetal[40] to investigate the relationship between the 146 candidate genes and the known disease genes, and found several GO functional categories (Figure 1) and pathways in which candidate genes were over-represented, and appeared to interact with other pathways that led to the pathogenesis of the disease. The 146 genes were enriched in signal transduction, positive regulation of cellular process, the immune system and immune

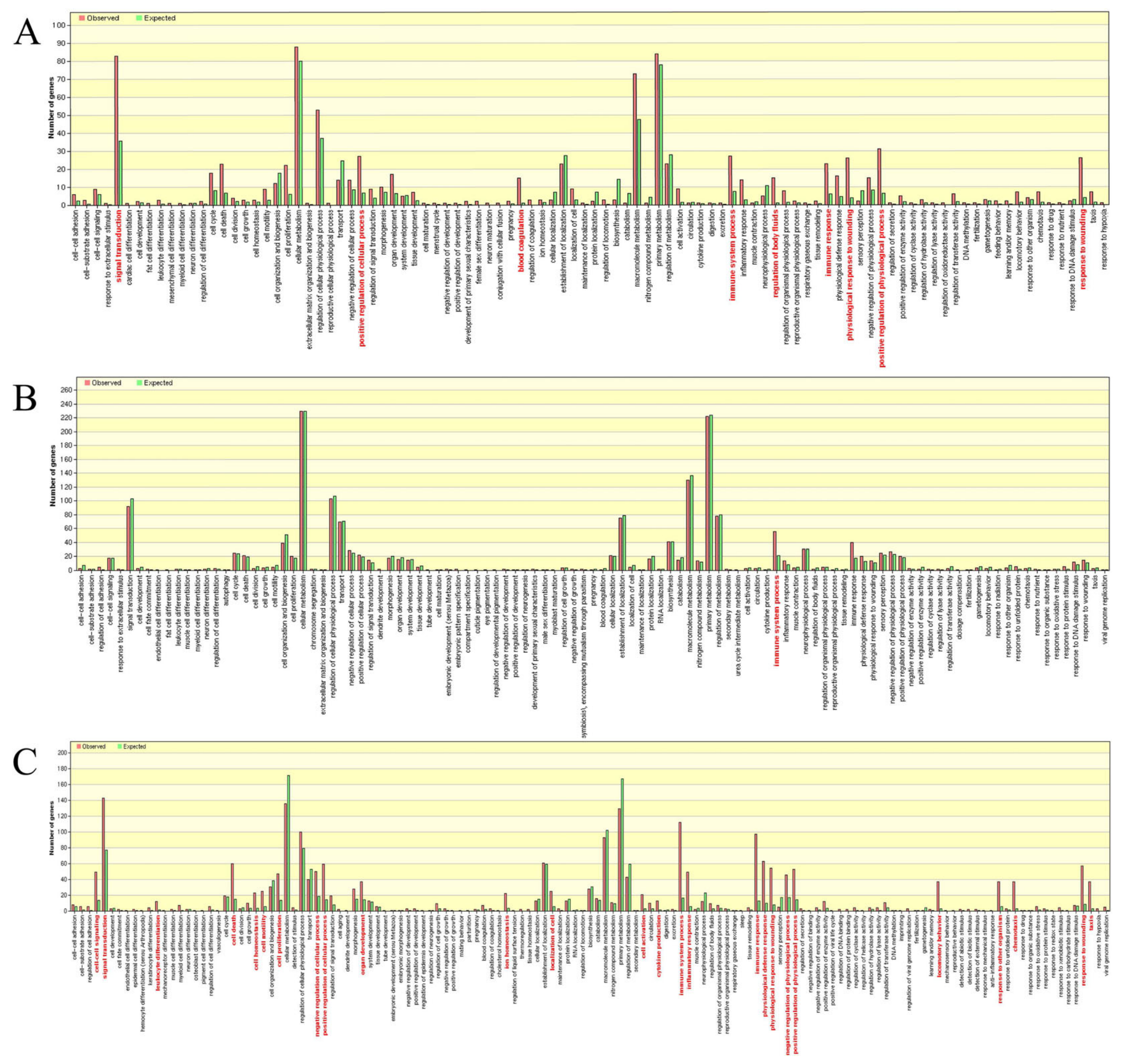

Figure 1 Results of GO functional enrichment of candidate genes and known disease genes. (A) GO functional enrichment of candidate genes predicted by our method. (B) GO functional enrichment of candidate genes predicted by the GWAS method. (C) GO functional enrichment of known disease genes.

response, and the physiological response to wounding, which was coincident with known disease genes (Figure 1). Responsible pathways included cytokine-cytokine receptor interaction pathways, Jak-STAT signaling pathways, cell adhesion molecules, and MAPK signaling pathways (Figure 2; Additional file 3 and 4). Candidate genes and known disease genes not only shared the same pathways, but also linked enriched pathways involved in passing on disease risk (Figure 2).

The nature of our screening approach meant that many of our predictions overlapped extensively in similar function categories. Therefore, to describe functions representative of association with rheumatoid arthritis, we selected those with the strongest association that also displayed a higher functional enrichment. For example, consistent with all previous studies of rheumatoid arthritis, genes in our gene set included members of the immunoglobulin protein family (Figure 3), the protein kinase domain family, the SH3 domain family, and the ligand-binding domains of nuclear hormone receptor family, and included several genes associated with moderate disease risk, and commonly reported genes such as CD4 [41-44], FGFR1 [45-47], and KDR [48-52]. Genes in the immunoglobulin protein family have a crucial role [53,54] in the pathogenesis of the disease. FGF-2 is transferred to FGFR-1 through binding to

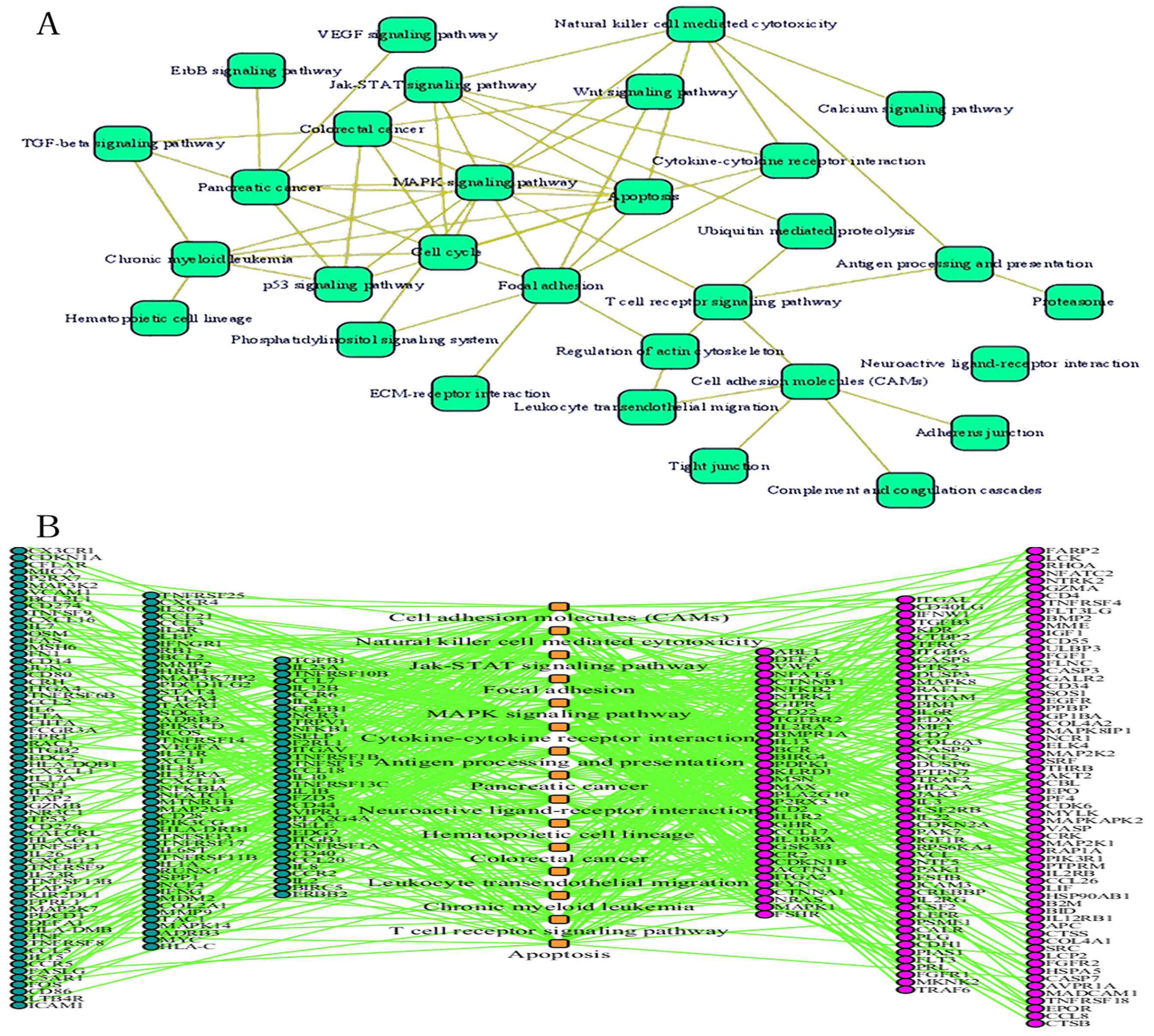

Figure 2 Pathway-pathway interactions and gene-pathway relationships. (A) Interactions between pathways of rheumatoid arthritis. Green vertices are responsible pathways and edges are interactions. (B) Relationships between candidate disease genes and corresponding pathways. Orange nodes are responsible pathways, blue nodes are known disease genes, and pink nodes are predicted disease genes.

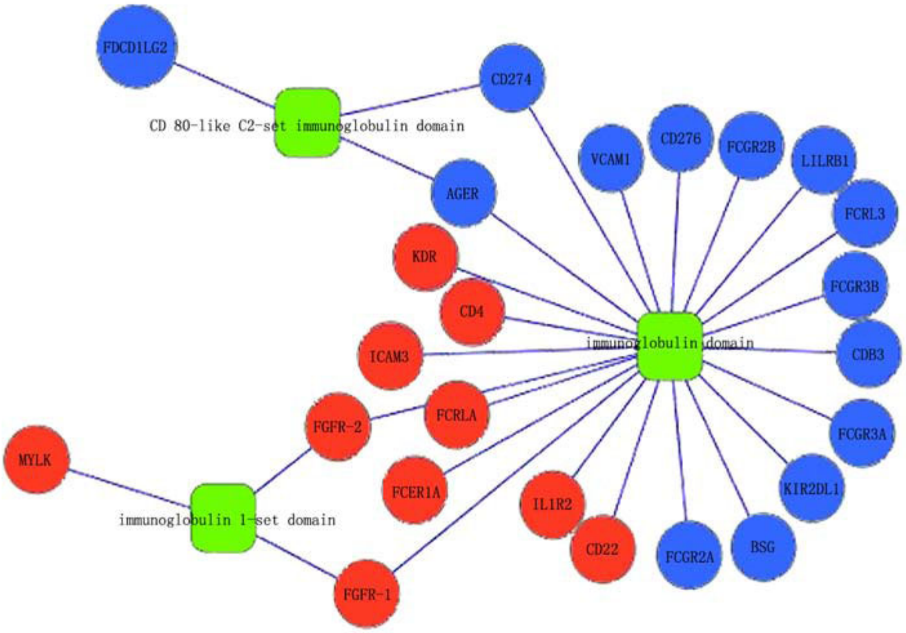

Figure 3 Known disease genes and predicted genes are enriched in immunoglobulin protein family. Nodes in blue represents known disease genes, red ones are predicted genes.

HSPG, resulting in RANKL and ICAM-1-mediated maturation of osteoclasts via ERK activation. FGF-2 not only augments the proliferation of RASFs, but is involved in osteoclast maturation, leading to bone destruction in rheumatoid arthritis. Genes in the immunoglobulin protein family also showed strong association with specific biological processes such as receptor binding, protein binding, and molecular transducer activity. Not surprisingly, a KEGG pathway search using these genes yielded the terms "cell adhesion molecules (CAMs)", "cytokine-cytokine receptor interaction" and "Jak-STAT signaling pathway" as the most significantly represented, similar to the enrichment analysis for known disease genes in these pathways. The candidate genes either interacted directly with enzymes associated with known disease genes, or conveyed disease risk indirectly.

We carried out GWAS to find the candidate gene set. The threshold of significant P-value (Bonferroni test) was set at 1.835e-8. GWAS identified 822 candidate genes (Additional file 5). We used Gene Webgetal software to check underlying biological associations for evidence of these candidates in the GO and KEGG databases during rheumatoid arthritis pathogenesis. Compared to traditional GWAS, which is based on multiple testing, we found that few candidate genes overlapped with the results of our multidimensional screening method. Most of the candidate genes in our prediction were verified as modestly associated with rheumatoid arthritis by literature retrieving, but were not identified by a traditional GWAs approach (Additional file 1). We note that a large number of candidate genes from the traditional prediction could not easily be classified into the related functional categories or interacting biological processes associated with this disease. This was not the case for our prediction, demonstrating the effectiveness of our proposed method (Figure 1, Additional file 3, 4 and 6). Candidate genes from GWAS tended to participate in immune systems processes (Figure 1), antigen processing and presentation, glutathione metabolism, cell adhesion molecules (CAMs) and glutathione metabolism and so on (Additional file 6). Even if dysfunction was found in these biological processes, little effect would be expected on other biological processes or pathways, and would not lead to systemic abnormalities or impairment in the function of human essential immune system (Figure 2A). Thus, we propose that the results from strictly statistical methods can find significant candidate genes, but does not consider minor- or medium-risk genes, and this might make uncovering the underlying pathogenesis of rheumatoid arthritis difficult for researchers in the post-genome area. Some candidates from our predicted results lack defined functional descriptions, and require further studies to verify their associations or mechanisms with rheumatoid arthritis, such as NTRK1, IL1R2, and SERPIND1.

Multidimensional approaches can also be applied to candidate gene identification of other diseases, where multiple genes share underlying biological similarities (*e. g.* the same pathway or GO term), or contribute to disease etiology but have common variations that make modest contributions to disease risk. Considering underlying biological similarities together with the proposed method, rather than focusing on a few SNPs or genes with the strongest evidence of disease association can detect likely causal genes. We hope that the proposed method provides additional insights into the pathogenesis of other diseases using hundreds of genetic variance in datasets.

Conclusions

In this article, we introduce a multi-dimensional screening approach to analyze the 16th Genetic Analysis Workshop (GAW16) data for rheumatoid arthritis, and identify candidate genes for rheumatoid arthritis. Our proposed approach is based on underlying biological similarities-based methods for candidate and known disease genes. Application of our method could identify likely candidate disease genes for rheumatoid arthritis, and could yield biological insights that are otherwise undetectable when focusing only on genes with the strongest evidence by multiple testing.

Traditional GWAS have been developed to identify susceptibility genes assuming a “most significant SNPs/genes” model. This screening process uses a strict selection of statistical thresholds, and aims to identify susceptibility genes based only on the statistical model, without considering multi-dimensional biological similarities in sequence arrangement, crystal structures, and functional categories or biological pathways shared between candidate and known disease genes. Thus, many minor or modestly associated risk genes are likely to be missed after multiple testing adjustments. GWAS and our methods have different objectives. The aim of our method is to avoid arbitrary multiple testing so that more risk biomarkers can be considered. Rather than focusing on individual genes for which evidence is strongest, our multidimensional screening approach typically extracts all risk SNPs/genes ($BFLn > 0$) by their odds ratios for hypothesis H_1 to H_0 , and looks for genes that share underlying biological similarities with known disease genes. We identified multiple genes sharing underlying biological similarities that contributed to disease etiology, but for which common variations made modest contributions to disease risk. A large number of candidate genes from traditional prediction could not be easily classified into related functional categories or interacting biological processes that are associated with the disease.

By considering underlying biological similarities together, rather than focusing on a few SNPs or genes with the strongest evidence of disease association, we can detect likely causal genes using the predicted method. We hope this alternative model complements the most significant SNPs/genes model, and provides additional insights into the pathogenesis of rheumatoid arthritis and other diseases, when using hundreds of genetic variance datasets.

Additional material

Additional file 1: known disease genes collected from the OMIM database and NCBI database.

Additional file 2: candidate genes with their corresponding risk SNPs predicted by our proposed method and biological evidences between genes and rheumatoid arthritis.

Additional file 3: the KEGG functional enrichment of candidate genes predicted by our proposed method.

Additional file 4: the KEGG functional enrichment of known disease genes.

Additional file 5: candidate genes with their corresponding risk SNPs out of GWAS and their corresponding risk SNPs and biological evidences between genes and rheumatoid arthritis.

Additional file 6: the KEGG functional enrichment of candidate genes out of GWAS.

Acknowledgements
The Genetic Analysis Workshops are supported by the NIH grant R01 GM031575 from the National Institute of General Medical Sciences. We thank the 16th Genomic Analysis Workshop (GAW16) for the generosity of providing us with the genotype frequency of SNP data.
Funds: National High Tech Development Project of China, the 863 Program (Grant No. 2007AA02Z329); the National Natural Science Foundation of China (Grant Nos. 30571034 and 30570424); the National Science Foundation of Heilongjiang Province (Grant Nos. D2007-48); Master Innovation Funds of Harbin Medical University (HCXS 2008010).

Authors’ contributions
LCZ and LNC guided the research and analyses described in the paper. LCZ and WL carried out multi-dimensional screening analysis, and LLS participated in performance evaluation of the results. LCZ and LNC participated in coordination of the study. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 3 December 2009 Accepted: 20 August 2010
Published: 20 August 2010

References

1. Dhaouadi T, Sfar I, Abelmoula L, Jendoubi-Ayed S, Aouadi H, Ben Abdellah T, Ayed K, Zouari R, Gorgi Y: **Role of immune system, apoptosis and angiogenesis in pathogenesis of rheumatoid arthritis and joint destruction, a systematic review.** *Tunis Med* 2007, **85(12)**:991-998.
2. Kanat F, Levendoglu F, Teke T: **Radiological and functional assessment of pulmonary involvement in the rheumatoid arthritis patients.** *Rheumatol Int* 2007, **27(5)**:459-466.
3. Yoo YJ, Gao G, Zhang K: **Case-control association analysis of rheumatoid arthritis with candidate genes using related cases.** *BMC Proc* 2007, **1(Suppl 1)**:S33.

4. Ritchie MD, Bartlett J, Bush WS, Edwards TL, Motsinger AA, Torstenson ES: **Exploring epistasis in candidate genes for rheumatoid arthritis.** *BMC Proc* 2007, **1(Suppl 1)**:S70.
5. Dieguez-Gonzalez R, Akar S, Calaza M, Gonzalez-Alvaro I, Fernandez-Gutierrez B, Lamas JR, de la Serna AR, Caliz R, Blanco FJ, Pascual-Salcedo D, *et al*: **Lack of Association with Rheumatoid Arthritis of Selected Polymorphisms in 4 Candidate Genes: CFH, CD209, Eotaxin-3, and MHC2TA.** *J Rheumatol* 2009.
6. Linsel-Nitschke P, Schunkert H, Erdmann J: **[Congestive heart failure is a common disease with complex inheritance–new perspectives through genome wide association studies].** *Internist (Berl)* 2008, **49(4)**:405-410, 412.
7. Evans DM, Visscher PM, Wray NR: **Harnessing the Information Contained Within Genome-wide Association Studies to Improve Individual Prediction of Complex Disease Risk.** *Hum Mol Genet* 2009.
8. Ohashi J, Tokunaga K: **The power of genome-wide association studies of complex disease genes: statistical limitations of indirect approaches using SNP markers.** *J Hum Genet* 2001, **46(8)**:478-482.
9. Wang K, Li M, Bucan M: **Pathway-Based Approaches for Analysis of Genomewide Association Studies.** *Am J Hum Genet* 2007, **81(6)**.
10. Chen L, Zhang L, Zhao Y, Xu L, Shang Y, Wang Q, Li W, Wang H, Li X: **Prioritizing risk pathways: a novel association approach to searching for disease pathways fusing SNPs and pathways.** *Bioinformatics* 2009, **25(2)**:237-242.
11. Waniek PJ, Castro HC, Sathler PC, Miceli L, Jansen AM, Araujo CA: **Two novel defensin-encoding genes of the Chagas disease vector Triatoma brasiliensis (Reduviidae, Triatominae): gene expression and peptide-structure modeling.** *J Insect Physiol* 2009, **55(9)**:840-848.
12. Stewart-Jones GB, Gillespie G, Overton IM, Kaul R, Roche P, McMichael AJ, Rowland-Jones S, Jones EY: **Structures of three HIV-1 HLA-B*5703-peptide complexes and identification of related HLAs potentially associated with long-term nonprogression.** *J Immunol* 2005, **175(4)**:2459-2468.
13. Wuritu , Ozawa Y, Gaowa , Kawamori F, Masuda T, Masuzawa T, Fujita H, Ohashi N: **Structural analysis of a p44/msp2 expression site of Anaplasma phagocytophilum in naturally infected ticks in Japan.** *J Med Microbiol* 2009, **58(Pt 12)**:1638-1644.
14. Yu C, Zavaljevski N, Stevens FJ, Yackovich K, Reifman J: **Classifying noisy protein sequence data: a case study of immunoglobulin light chains.** *Bioinformatics* 2005, **21(Suppl 1)**:i495-501.
15. Melvin I, Weston J, Leslie CS, Noble WS: **Combining classifiers for improved classification of proteins from sequence or structure.** *BMC Bioinformatics* 2008, **9**:389.
16. Joung JG, Fei Z: **Computational identification of condition-specific miRNA targets based on gene expression profiles and sequence information.** *BMC Bioinformatics* 2009, **10(Suppl 1)**:S34.
17. Kuo CH, Miyazaki D, Nawata N, Tominaga T, Yamasaki A, Sasaki Y, Inoue Y: **Prognosis-determinant candidate genes identified by whole genome scanning in eyes with pterygia.** *Invest Ophthalmol Vis Sci* 2007, **48(8)**:3566-3575.
18. Hertel J, Hofacker IL, Stadler PF: **SnoReport: computational identification of snoRNAs with unknown targets.** *Bioinformatics* 2008, **24(2)**:158-164.
19. Hong MG, Pawitan Y, Magnusson PK, Prince JA: **Strategies and issues in the detection of pathway enrichment in genome-wide association studies.** *Hum Genet* 2009, **126(2)**:289-301.
20. Li QX, Ke N, Sundaram R, Wong-Staal F: **NR4A1, 2, 3–an orphan nuclear hormone receptor family involved in cell apoptosis and carcinogenesis.** *Histol Histopathol* 2006, **21(5)**:533-540.
21. Stuffers S, Brech A, Stenmark H: **ESCRT proteins in physiology and disease.** *Exp Cell Res* 2009, **315(9)**:1619-1626.
22. Rusten TE, Filimonenko M, Rodahl LM, Stenmark H, Simonsen A: **ESCRTing autophagic clearance of aggregating proteins.** *Autophagy* 2007, **4(2)**.
23. Lebrec JJ, Huizinga TW, Toes RE, Houwing-Duistermaat JJ, van Houwelingen HC: **Integration of gene ontology pathways with North American Rheumatoid Arthritis Consortium genome-wide association data via linear modeling.** *BMC Proc* 2009, **3(Suppl 7)**:S94.
24. Arya R, Hare E, Del Rincon I, Jenkinson CP, Duggirala R, Almasy L, Escalante A: **Effects of covariates and interactions on a genome-wide association analysis of rheumatoid arthritis.** *BMC Proc* 2009, **3(Suppl 7)**:S84.
25. Kang G, Childers DK, Liu N, Zhang K, Gao G: **Genome-wide association studies of rheumatoid arthritis data via multiple hypothesis testing methods for correlated tests.** *BMC Proc* 2009, **3(Suppl 7)**:S38.
26. Koffeman EC, Genovese M, Amox D, Keogh E, Santana E, Matteson EL, Kavanaugh A, Molitor JA, Schiff MH, Posever JO, *et al*: **Epitope-specific immunotherapy of rheumatoid arthritis: clinical responsiveness occurs with immune deviation and relies on the expression of a cluster of molecules associated with T cell tolerance in a double-blind, placebo-controlled, pilot phase II trial.** *Arthritis Rheum* 2009, **60(11)**:3207-3216.
27. van der Pouw Kraan TC, van Baarsen LG, Wijbrandts CA, Voskuyl AE, Rustenburg F, Baggen JM, Dijkmans BA, Tak PP, Verweij CL: **Expression of a pathogen-response program in peripheral blood cells defines a subgroup of rheumatoid arthritis patients.** *Genes Immun* 2008, **9(1)**:16-22.
28. Jawaheer D, Lum RF, Amos CI, Gregersen PK, Criswell LA: **Clustering of disease features within 512 multicase rheumatoid arthritis families.** *Arthritis Rheum* 2004, **50(3)**:736-741.
29. Dotzlaw H, Schulz M, Eggert M, Neeck G: **A pattern of protein expression in peripheral blood mononuclear cells distinguishes rheumatoid arthritis patients from healthy individuals.** *Biochim Biophys Acta* 2004, **1696(1)**:121-129.
30. Liu W, Li X, Ding F, Li Y: **Using SELDI-TOF MS to identify serum biomarkers of rheumatoid arthritis.** *Scand J Rheumatol* 2008, **37(2)**:94-102.
31. Barton A, Thomson W, Ke X, Eyre S, Hinks A, Bowes J, Gibbons L, Plant D, Wilson AG, Marinou I, *et al*: **Re-evaluation of putative rheumatoid arthritis susceptibility genes in the post-genome wide association study era and hypothesis of a key pathway underlying susceptibility.** *Hum Mol Genet* 2008, **17(15)**:2274-2279.
32. Waldron ER, Whittaker JC, Balding DJ: **Fine mapping of disease genes via haplotype clustering.** *Genet Epidemiol* 2006, **30(2)**:170-179.
33. Alibes A, Yankilevich P, Canada A, Diaz-Uriarte R: **IDconverter and IDClight: conversion and annotation of gene and protein IDs.** *BMC Bioinformatics* 2007, **8**:9.
34. Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ: **SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence.** *Nucleic Acids Res* 2003, **31(13)**:3692-3697.
35. Pirooznia M, Deng Y: **SVM Classifier - a comprehensive java interface for support vector machine classification of microarray data.** *BMC Bioinformatics* 2006, **7(Suppl 4)**:S25.
36. Sato JR, da Graca Morais Martin M, Fujita A, Mourao-Miranda J, Brammer MJ, Amaro E Jr: **An fMRI normative database for connectivity networks using one-class support vector machines.** *Hum Brain Mapp* 2009, **30(4)**:1068-1076.
37. Du R, Tantisira K, Carey V, Bhattacharya S, Metje S, Kho AT, Klanderman BJ, Gaedigk R, Lazarus R, Mariani TJ, *et al*: **Platform dependence of inference on gene-wise and gene-set involvement in human lung development.** *BMC Bioinformatics* 2009, **10**:189.
38. Takeda S: **Three-dimensional domain architecture of the ADAM family proteinases.** *Semin Cell Dev Biol* 2009, **20(2)**:146-152.
39. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics* 2005, **21(2)**:263-265.
40. Zhang B, Kirov S, Snoddy J: **WebGestalt: an integrated system for exploring gene sets in various biological contexts.** *Nucleic Acids Res* 2005, , **33 Web Server:** W741-748.
41. Olteanu H, Karandikar NJ, Eshoa C, Kroft SH: **Laboratory findings in CD4(+) large granular lymphocytoses.** *Int J Lab Hematol* **32(1 Pt 1)**:e9-16.
42. Suzuki K, Saito K, Tsujimura S, Nakayamada S, Yamaoka K, Sawamukai N, Iwata S, Nawata M, Nakano K, Tanaka Y: **Tacrolimus, a calcineurin inhibitor, overcomes treatment unresponsiveness mediated by P-glycoprotein on lymphocytes in refractory rheumatoid arthritis.** *J Rheumatol* **37(3)**:512-520.
43. Swainson LA, Mold JE, Bajpai UD, McCune JM: **Expression of the Autoimmune Susceptibility Gene FcRL3 on Human Regulatory T Cells Is Associated with Dysfunction and High Levels of Programmed Cell Death-1.** *J Immunol* .
44. Tukaj S, Kotlarz A, Jozwik A, Smolenska Z, Bryl E, Witkowski JM, Lipinska B: **Hsp40 proteins modulate humoral and cellular immune response in rheumatoid arthritis patients.** *Cell Stress Chaperones* .
45. Yamane S, Ishida S, Hanamoto Y, Kumagai K, Masuda R, Tanaka K, Shiobara N, Yamane N, Mori T, Juji T, *et al*: **Proinflammatory role of amphiregulin, an epidermal growth factor family member whose expression is augmented in rheumatoid arthritis patients.** *J Inflamm (Lond)* 2008, **5**:5.
46. Green N, Hu Y, Janz K, Li HQ, Kaila N, Guler S, Thomason J, Joseph-McCarthy D, Tam SY, Hotchandani R, *et al*: **Inhibitors of tumor progression loci-2 (Tpl2) kinase and tumor necrosis factor alpha (TNF-alpha)**

production: selectivity and in vivo antiinflammatory activity of novel 8-substituted-4-anilino-6-aminoquinoline-3-carbonitriles. *J Med Chem* 2007, **50(19)**:4728-4745.

47. Hallbeck AL, Walz TM, Briheim K, Wasteson A: **TGF-alpha and ErbB2 production in synovial joint tissue: increased expression in arthritic joints.** *Scand J Rheumatol* 2005, **34(3)**:204-211.
48. Yiu KH, Wang S, Mok MY, Ooi GC, Khong PL, Lau CP, Lai WH, Wong LY, Lam KF, Lau CS, *et al*: **Role of circulating endothelial progenitor cells in patients with rheumatoid arthritis with coronary calcification.** *J Rheumatol* **37(3)**:529-535.
49. Ablin JN, Goldstein Z, Aloush V, Matz H, Elkayam O, Caspi D, Swartzenberg S, George J, Wohl Y: **Normal levels and function of endothelial progenitor cells in patients with psoriatic arthritis.** *Rheumatol Int* 2009, **29(3)**:257-262.
50. Egan CG, Caporali F, Garcia-Gonzalez E, Galeazzi M, Sorrentino V: **Endothelial progenitor cells and colony-forming units in rheumatoid arthritis: association with clinical characteristics.** *Rheumatology (Oxford)* 2008, **47(10)**:1484-1488.
51. Lazarovici P, Marcinkiewicz C, Lelkes PI: **Cross talk between the cardiovascular and nervous systems: neurotrophic effects of vascular endothelial growth factor (VEGF) and angiogenic effects of nerve growth factor (NGF)-implications in drug development.** *Curr Pharm Des* 2006, **12(21)**:2609-2622.
52. Grisar J, Aletaha D, Steiner CW, Kapral T, Steiner S, Seidinger D, Weigel G, Schwarzinger I, Wolozcszuk W, Steiner G, *et al*: **Depletion of endothelial progenitor cells in the peripheral blood of patients with rheumatoid arthritis.** *Circulation* 2005, **111(2)**:204-211.
53. Bohnhorst JO, Hanssen I, Moen T: **Immune-mediated fever in the dog. Occurrence of antinuclear antibodies, rheumatoid factor, tumor necrosis factor and interleukin-6 in serum.** *Acta Vet Scand* 2002, **43(3)**:165-171.
54. Nakano K, Okada Y, Saito K, Tanaka Y: **Induction of RANKL expression and osteoclast maturation by the binding of fibroblast growth factor 2 to heparan sulfate proteoglycan on rheumatoid synovial fibroblasts.** *Arthritis Rheum* 2004, **50(8)**:2450-2458.

Pre-publication history

The pre-publication history for this paper can be accessed here:
http://www.biomedcentral.com/1755-8794/3/38/prepub

doi:10.1186/1755-8794-3-38
Cite this article as: Zhang *et al.*: A towards-multidimensional screening approach to predict candidate genes of rheumatoid arthritis based on SNP, structural and functional annotations. *BMC Medical Genomics* 2010 **3**:38.