Author's response to reviews

Title: A Towards-Multidimensional Screening Approach to Predict Candidate Genes of Rheumatoid Arthritis based on SNP, Structural and Functional Annotations

Authors:

   Zhang Liang Cai (zhangliangcai@ems.hrbmu.edu.cn)
   Lina Chen (chenglina_2004@yahoo.com.cn)

Version: 3 Date: 14 April 2010

Author's response to reviews: see over
Mar 27, 2010

The Editor(s) of BioMed Central Editorial
Email: editorial@biomedcentral.com

RE: submission of the revised version for the article (Ms id# 1575330671331404) entitled with “A Towards-Multidimensional Screening Approach to Predict Candidate Genes of Rheumatoid Arthritis based on SNP, Structural and Functional Annotations” Zhang and Chen et al.

Dear Editor(s) for BioMed Central:

Thank you for your editorial efforts for our manuscript, and your recommendation that we submit a revision of this manuscript. We also thank much the anonymous reviewer for his (her) constructive comment to strengthen this manuscript, which has been accommodated appropriately.

We have carefully revised the manuscript to respond to the points and concerns raised by the reviewers. In particular, we have undertaken the following major revisions over the previous version:

1) Per the reviewer’s advice “More explanation of their approach will be necessary for better understanding”, we have carefully given more explanation of our approach to make it better understood. We have explained that why we choose Bayes Factor and SVM method and why a Bayesian screening is a necessary step before SVM method. We have reinforced the description in the sections of ‘Abstract’ and ‘Background’ in the revised paper.

2) Regarding the reviewer’s advice that the SVM screening is not clear based on it is written; we thank for the reviewer’s careful comments, and have revised the error description of SVM. And after that, we have added essential steps and the corresponding results that we carried out during the second filtering. The corresponding revision could be found in the sections of ‘Result and Discussion’ and ‘Method’.

3) To respond to “the notation for Beta function and several same questions for the numbers of cases and controls” of the reviewer, we have revised the corresponding notations to be correct following the reviewer’s advice. And after that, we made the notation to be consistent in the remainder of the paper.

4) To respond to the method comparison issue by the reviewer, we have accommodated this section by discussing the results of our proposed method and the results of GWAS to illustrate the efficiency of our method (see in supplementary table 1-6 and supplementary figure 1-3).

5) Per the reviewer’s advice “The specific research question posed by the authors is not well defined” by the reviewer, we have given adequate description and explanation for defining our research question in the following reply and in the corresponding abstract section.
In addition to accommodating the comments and concerns of the reviewers, we have carefully corrected the descriptive errors in the previous version to improve its overall clarity and readability. Some minor mistakes have been rectified per the reviewer’s advice. The manuscript has been internally reviewed and proofread by several professionals and the commercial copyediting service, International Science Editing.

In the revised manuscript, major revisions are shown in **BLUE** color. The detailed responses on a point-by-point basis and revisions made are described below.

We believe that the manuscript has been greatly strengthened by the critique of the reviewer and hope that both you and the reviewers will now find the paper suitable for publication.

Thank you and the reviewers again for carefully reviewing our work.

Respectfully yours,

Lina Chen, Ph.D.
College of Bioinformatics Science and Technology, Harbin Medical University, China.
Phone: (+086)0451-86674768
Email: chenlina@ems.hrbmu.edu.cn and zhangliangcai@ems.hrbmu.edu.cn
Reply to Reviewer #1 (report from 2 February 2010, Reviewer: Euijung Ryu)

(1). “The authors introduced a new method called towards-multidimension screening approach. More explanation of their approach will be necessary for better understanding. For example, why do they choose Bayes Factor and SVM method? Screening the SNPs using Bayes Factor was necessary step for SVM method?”

REPLY: Thank the reviewer for identifying the insufficient description in the explanation of our approach. We accepted the reviewer’s advice that these details should be stated. Hundreds of thousands of SNPs have been tested for rheumatoid arthritis according to the Genetic Analysis Workshops (GAWs). As we all know, traditional genome-wide association studies (GWAS) have been developed to identify susceptibility genes assuming a “most significant SNPs/genes” model. And meanwhile, lots of risk genes (e.g. minor or modest) are likely to be missed after adjustment of multiple testing. This screening process resorts to a strict selection of statistical thresholds, which aims to identify susceptibility genes based only on the statistic model without considering multi-dimensional biological similarities in sequence arrangement, crystal structures and functional categories/biological pathways between candidate and known disease genes.

The objectives are different between GWAS and our Bayesian screening analyses. We have carefully read the corresponding traditional genome-wide association studies and its supplementary information. The aim of traditional genome-wide association approaches is to identify most significant biomarkers and often miss risk biomarkers that not fit the strict statistical threshold of multiple testing, making their screening difficult to interpretation. However, the idea and rationale behind Bayesian approaches is to compare the probability if there is an association to its probability given no association at a single SNP site. The aim of genome-wide Bayesian association analysis is to avoid arbitrary multiple testing so that more risk biomarkers could be taken into account. We hope this alternative model could complement for the “most significant SNPs/genes” model to test for association with case-control phenotype of each SNP. And moreover, Bayesian association tests are often considered as either equal or superior to classical tests in terms of objectivity, ease of use, and ease of interpretation (GREG M. ALLENBY. Hypothesis testing with scanner data: the advantage of Bayesian methods). Comparing with GWAS, We have carried out the SNP significant analysis using Bayesian association approach, and found that 273(86.4%) risk SNPs out of GWAS are included in our Bayesian screening result set.

Support vector machines (SVM) is a machine learning algorithm based on the Statistical Learning Theory (SLT), and is commonly applied to resolve the classification problems, which can get good classification effects with a few learning samples. In this paper, we hypothesize that underlying candidate genes harboring markers with minor or modest evidence of association could be identified if they share similar attributions with known disease genes in underlying biological annotations such as the same sequence arrangement of their corresponding genes, crystal structure of encoding proteins and the same biological pathways or mechanisms. During the second stage of our multi-dimensional screening, SVM method has been used to distinguish likely-causing genes from non-disease genes according to different sequential and crystal structural features of their proteins.

In our analysis, the SNP screening using Bayes Factor is a necessary step for the second screening (SVM method): i) our multi-dimensional screening approach is designed to identify susceptible genes for population genetic variance datasets in statistical genetics area. Bayesian association analysis is also called genetic screening to get risk biomarkers for the given disease. ii) After the SNP screening using Bayes Factor, a set of risk genes could be labeled as disease risk genes. And then a negative gene set for the second SVM screening could be obtained after exception of known disease genes and these disease risk genes.

We have revised the corresponding sections overall according to reviewer’s advice and have incorporated the following elaborations in the revised manuscript:
In the section of ‘Abstract’:

“Abstract Background: Hundreds of thousands of SNPs have been tested for rheumatoid arthritis according to the Genetic Analysis Workshops (GAWs). Traditional genome-wide association studies (GWAS) have been developed to identify susceptibility genes assuming a “most significant SNPs/genes” model. And meanwhile, lots of minor or modest risk genes are likely to be missed after adjustment of multiple testing, and this screening process resorts to a strict selection of statistical thresholds, which aims to identify susceptibility genes based only on the statistic model without considering multi-dimensional biological similarities in sequence arrangement, crystal structures and functional categories/biological pathways between candidate and known disease genes. Multidimensional screening approaches, combined with traditional statistical genetics methods, could consider multiple biological backgrounds of genetic mutation, structural and functional annotations. Here we introduce a towards-multidimensional screening approach to predicate candidate genes of rheumatoid arthritis that take into account all SNPs with nominal evidence of Bayesian association (BFLn>0), structural and functional similarities of their corresponding genes or proteins.

Results: Traditional candidate gene prediction studies return 822 genes and a large part of these candidate genes could not easily be classified into related functional categories or biological processes associated to this disease, while our prediction could prevent this phenomenon occurring. Finally, we identified 146 candidate genes of rheumatoid arthritis including gene CD4, FGFR1 and KDR and so on, which were reported as high risk factors by most recent studies. And meanwhile, some minor or modest disease risk genes could be identified by our proposed method, and the associations between these genes and this disease are verified by literature retrieving. And it could identify candidate likely-causing disease genes of rheumatoid arthritis. We hope that using the proposed method could provide additional insights into the pathogenesis of rheumatoid arthritis and other diseases, when handling tens of hundreds of genetic variances datasets.”

In the section of “Background”:

“When case–control datasets of complex diseases are available, …..

….. are likely to be missed in the ‘most significant SNPs/genes’ approach [9,10]. And moreover, this screening process aims to identify susceptibility genes based only on the statistic model without considering essential biological background of genetic mutation, structural and functional annotations. To avoid strict adjustment of multiple testing, we introduce a genome-wide Bayesian association method to test for association with case-control phenotype at a single SNP. The idea behind Bayesian approaches is to compare the probability if there is an association to its probability given no association. Based on genome-wide Bayesian association methods, discovering new bioinformatics strategies to avoid the limitations of studies of complex diseases is vital.

Traditional statistic genetics aims to identify susceptibility genes based only on the statistic model without considering biological similarities between disease genes and likely-causing genes. As we all know, proteins are essential parts of organisms and participate in virtually every process within cells and most proteins fold into unique sequence arrangements and structures and contribute to specific characteristics in diverse function sets. Responsible proteins (genes) of complex diseases are often associated with sort of similar sequences and structures [11-13], and researchers could carry on screening candidate genes according to similarities in their sequence arrangement and crystal structure shared with known disease genes. Support vector machines (SVM) is a machine learning algorithm based on the Statistical Learning Theory (SLT), and is commonly applied to resolve this problem [14-18], which can get good classification effects with a few learning samples.”

(2). “Do the authors used "trainig set" for test set, which was for candidate genes? Or, is it part of cross-validation? It is not clear based on how it is written.”

REPLY: Thank the reviewer for pointing it out. During the SVM screening, the positive set consists of disease genes of rheumatoid arthritis gathered from OMIM online database. The test set consists of risk genes out of Bayesian association screening. While, the negative set contains the remaining genes that won’t fall within any disease loci after excluding genes in the positive set and the test set. A 28-dimension vector of physicochemical features, a combinational pseudo-sequence, is used to represent each protein in positive, negative and testing sets according to the online RCSB PDB and targetDB databases. The second screening process (SVM classification based on sequence and structure similarity feature) of our method can be divided into two parts: feature selection and candidate gene prediction. (Part I) Feature selection: Here,
8-dimension secondary feature (21-28) and the entire 28-dimension physicochemical feature are used to train two classifiers and we carry out the second screening. 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) was used for training. After 1,000 randomization, we found the classifier of 28-dimension physicochemical feature is more accurate. (Part II) 

**Candidate gene prediction:** We utilize combined 28-dimension physicochemical feature, the positive set, the negative set and the test set to retrain the classifier. We construct 10,000 additional training sets (Positive/Negative=335:335), in which each negative set is selected randomly from the original negative set. We perform this process for 10,000 times and keep the genes which are judged to be disease genes in each process of randomization as our predictions. And we have rewritten the corresponding description of methods overall in the Method section according to reviewer’s advice. We have incorporated the following elaborations in the revised paragraphs:

In the section of ‘Method’:

“**SVM Classification based on Sequence and Structure Similarity Features**

Firstly, ID Converter[48] is used to map all the genes to their corresponding proteins across the whole human genome. In this section, the positive set consists of disease genes of rheumatoid arthritis gathered from NCBI and OMIM online database (see in supplementary table 3). While, the negative set contains the remaining genes that won’t 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 (in table 3), a combinational pseudo-sequence, is used to represent each protein in positive, negative and testing sets according to the online RCSB PDB and targetDB databases. Here, 8-dimension secondary features (21-28) and the entire 28-dimension physicochemical features are used to train two classifiers and we carry out the second screening.

Considering the diversity of the putative non disease-candidate proteins, the non disease-candidate space might not have been sampled completely. Therefore, we construct 1,000 additional training sets (Positive/Negative=1:1), in which each negative set is selected randomly from the original negative set. During each randomization, the 8- and 28-dimension feature are used to construct the corresponding classifier. The performance of our model is evaluated with an n-fold cross-validation test. In the cross-validation test, the entire positive and negative data sets are shuffled and split into n folds. Each fold is used in turn for testing and the remaining part (n-1 folds) is used for training. The sensitivity ($Q_p$), specificity ($Q_n$) and overall accuracy ($Q_a$) are used to measure the accuracy of positive prediction, negative prediction and the overall accuracy of the model [49], respectively.

\[
Q_p = \frac{TP}{TP + FN} \\
Q_n = \frac{TN}{TN + FP} \\
Q_a = \frac{(TP + TN)}{(TP + TN + FP + FN)}
\]

Here, TP, TN, FP and FN represent true positives, true negatives, false positives and false negatives, respectively. In general, the overall accuracy $Q_a$ is always used to measure the predictive power of a model.

We construct 10,000 additional training sets (Positive/Negative=335:335), in which each negative set is selected randomly from the original negative set. Here, the test set is prepared from genetic screenings. We utilize the training sets, the test set and the optima classifier to retrain the classifier. In this process, we aim to classify the genes in the test set, and then predict candidate genes of RA. We perform randomization 10,000 times and keep the genes which are judged to be disease genes in each process of random as our predictions.”

(3). “In page 5, the authors mentioned "well-filtered pair of positive and negative sample sets". What is the criterion for filtering?”

**REPLY:** Thank the reviewer very much for pointing out the error description in page 5. In fact, during this process, we have carried out the following procedures. After the feature selection, we construct 10,000 additional training sets (Positive/Negative=335:335), in which each negative set is selected randomly from the original negative set. Here, the test set is prepared from genetic screenings. We utilize the training sets, the test set and the optima classifier to retrain the classifier. In this process, we aim to classify the genes in the test set, and then predict candidate genes of RA. We perform randomization 10,000 times and keep the genes which are judged to be disease genes in each process as our predictions. According to the reviewer’s advice, we have reorganized
the SVM screening part in the method section and revised the improper descriptions in the corresponding result section. We have incorporated the following elaborations in the revised paragraphs:

In the section of ‘Results and Discussion:

“…… This process resulted in 4,402 risk candidate genes, which were labeled as members of the test set for the SVM screening step.

To carry out the SVM screening analysis, we extracted the sequence and structure information from PDB and targetDB databases and calculated their feature values for 335 known disease genes, 28,874 non-disease genes and other 4,402 genes in the test set, respectively. Here, 8-dimension secondary features (21-28) and the entire 28-dimension physicochemical features were used to train two classifiers and we carry out the second screening (see in method section). Considering the diversity of the putative non disease-candidate proteins, the non disease-candidate space might not have been sampled completely. Therefore, we constructed 1,000 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 classifiers. The performance of our model was evaluated with a 5-fold cross-validation test. In the cross-validation test, the entire positive and negative data sets were shuffled and split into 5 folds. Each fold was used in turn for testing and the remaining part (3-1 folds) was used for training. The 1,000 randomization results of two classifiers were carried out and the relevant accuracy of 28-dimension physicochemical features varied between 0.695 and 0.891 (Table 1).

Table 1 the performance information of two classifiers based on 8-dimension secondary physicochemical features and 28-dimension physicochemical features

<table>
<thead>
<tr>
<th>features</th>
<th>prediction</th>
<th>average ± std.</th>
</tr>
</thead>
<tbody>
<tr>
<td>8-dimension Secondary features(21-28)</td>
<td>0.631~0.831</td>
<td>0.703±0.038</td>
</tr>
<tr>
<td>28-dimension physicochemical features</td>
<td>0.695~0.891</td>
<td>0.762±0.036</td>
</tr>
</tbody>
</table>

According to the prediction of these two classifiers, we chose the second classifier and carried on candidate genes prediction. We reconstructed 10,000 additional training sets (Positive/Negative=335:335), in which each negative set is selected randomly from the original negative set. Here, the test set was prepared from genetic screenings. After 10,000 times randomization, the intersection of each prediction was defined as our final prediction. This process resulted in 495 candidate genes and they were used for the third screening step.”

(4). “In page 6, the authros listed several genes by referring to Figure 3. However, Figure 3 does not have names of genes. Instead it has some numbers in nodes. Are these numbers supposed to be corresponding to a certain gene? For example, which one is for CD4 gene?”

REPLY: Thank the reviewer very much for pointing out the insufficient description. Those numbers in the primary figure 3 are gene Ids. Per the reviewer’s advice, we have listed detailed descriptions of gene symbols instead of their corresponding Ids. We have incorporated the following elaborations in the revised figure 2:

In the section of ‘Results and Discussion:

“Due to the nature of our screening approach, many of our predictions overlap extensively in sorts of similar function categories. Thus, to describe function representative of association with rheumatoid arthritis, we selected those with the strongest association which also displayed a higher functional enrichment. For example, consistent with all the previous associated studies of rheumatoid arthritis, genes in our resulting gene set were members of the immunoglobulin protein family (see in figure 2), protein kinase domain family, SH3 domain family and ligand-binding domain of nuclear hormone receptor family and so on, and they included several moderate disease risk and hotly-reported genes such as CD4\cite{32-35}, FGFR1\cite{36-38} and KDR\cite{39-43}. These genes in the immunoglobulin protein family played a crucial role \cite{44, 45} in the pathogenesis of this disease:”

“Figure 2 - Known disease genes and predicted genes are enriched in immunoglobulin protein family. Nodes in blue represent known disease genes, red ones are predicted genes.”
(5). “In Figure 5, the result from GWAS is not concordant with one from the proposed method, as the authors mentioned. Does this mean that the genes selected from the proposed method (98 genes, for example) were too minor to be detected using GWAS? If so, what about 790 genes selected from GWAS? Why they are not chosen from their proposed method?”

REPLY: Thank the reviewer to point it out. Our proposed approach to the analysis of GAW data is motivated by underlying biological similarities-based methods between candidate and known disease genes. When genetic profiles of different variances are compared, tens or hundreds of variances (SNPs) may have subtle differences in case-control phenotypes. Rather than focusing on individual genes that have the strongest evidence, our multi-dimensional screening approaches typically extract all risk genes (BFLn>0) by their odd ratios of hypothesis $H_1$ to $H_0$ and then look for whether a particular group of genes shares underlying biological similarities with known disease genes. Application of our method has yielded biological insights that are otherwise undetectable by focusing only on genes with the strongest evidence of multiple testing. A large part of candidate genes out of traditional prediction could not easily be classified into related function categories or interacting biological processes associated with this disease (supplementary figure 2 and table 6). For example, when 790 genes were carried out functional enrichment studies on GO functions and pathway enrichments, these genes were prone to participate majorly in antigen processing and presentation, cell adhesion molecules (CAMs), glutathione metabolism and cytokine-cytokine receptor interaction and so on, even if dysfunction was found in any of these biological processes, it seems to have little effect on other processes and could not lead to systemically abnormality or impairment in the function of human essential immune system. And meanwhile, 98 genes out of our proposed method were too minor to be detected using GWAS and showed a significant and consistent state with known disease genes at function levels (supplementary figure 1 and table 4). We propose that multi-dimensional approaches can also be applied to candidate gene identification of rheumatoid arthritis, where multiple genes sharing underlying biological similarities (e.g. in the same pathway or GO term) contribute to disease etiology but where common variations in each of these genes make modest contributions to disease risk. Considering underlying biological similarities together, rather than focusing on a few SNPs and/or genes with the strongest evidence of disease association, we can detect likely-causing genes using the proposed method. We have revised the corresponding section to make it readable. We have incorporated the following elaborations in the revised paragraphs:

In the section of ‘VERIFICATION’:

“We carried out GWA studies to find the candidate gene set for this disease. The threshold of significant P-value (bonferroni test) was set at 1.835e-8. 822 candidate genes were identified by genome-wide association studies (see in supplementary table 2). Furthermore, we used the Gene Webgetal software to check underlying biological associations about evidences of these candidates in GO and KEGG databases during the pathogenesis of rheumatoid arthritis. Comparing with traditional genome-wide association studies, which were based on multiple testing, we found few candidate genes overlap with the results of our multi-dimensional screening method. And most of other candidate genes in our prediction were verified to be modestly-associated with rheumatoid arthritis by literature retrieving, but couldn’t be identified by traditional GWAs approach (see in supplementary table 1). And we must note that a large part of candidate genes out of traditional prediction could not easily be classified into all the related function categories or interacting biological processes associated with this disease, while our prediction could prevent this phenomenon occurring, which demonstrated the effectiveness of our proposed method (supplementary figure 1, 2 and 3; supplementary table 4, 5 and 6). Candidate genes out of GWA studies
were prone to participate majorly in the immune systems process (see in supplementary figure 3), antigen processing and presentation, cell adhesion molecules (CAMs) and glutathione metabolism and so on (see in supplementary table 6), even if dysfunction was found in any of these biological processes, and it seemed to have little effect on other biological processes or pathways and could not lead to systemically abnormality or impairment in the function of human essential immune system (figure 1A). Thus, we argued that these results from strictly-statistic based methods could find significant candidate genes, but its lack consideration of minor or medium risk genes might make it difficult to uncover the underlying pathogenesis of rheumatoid arthritis for researchers in the post-genome area. Some candidates of our predicted results, lack of defined function descriptions, might need further studies to verify their inner associations or mechanisms with RA, such as NTRK1, IL1R2 and SERPIND1 and so on.

Multi-dimensional approaches can also be applied to candidate gene identification of other diseases, where multiple genes sharing underlying biological similarities (e.g. in the same pathway or GO term) contribute to disease etiology but where common variations in each of these genes make modest contributions to disease risk. Considering underlying biological similarities together, rather than focusing on a few SNPs and/or genes with the strongest evidence of disease association, researches can detect likely-causing genes using the proposed method. We hope that the proposed method could provide additional insights into the pathogenesis of other diseases, when handling tens of hundreds of genetic variances datasets.”

In the section of supplemental file:

“Supplementary figure 1 - the GO functional enrichment of candidate genes predicted by our proposed method.tif”
“Supplementary figure 2 - the GO functional enrichment of candidate genes out of GWAS.tif”
“Supplementary figure 3 - the GO functional enrichment of known disease genes.tif”
“Supplementary table 4 - the KEGG functional enrichment of candidate genes predicted by our method.xls”
“Supplementary table 5 - the KEGG functional enrichment of known disease genes.xls”
“Supplementary table 6 - the KEGG functional enrichment of candidate genes out of GWAS.xls”

(6). “What is the basis to claim that the proposed method is more precise than GWAS? What measure was used to claim it?”

REPLY: Thank the reviewer to point it out. We accept the reviewer’s advice and better explain this issue. Traditional candidate gene prediction studies return 822 genes and a large part of these candidate genes could not easily be classified into related functional categories or biological processes associated to this disease, while our prediction could prevent this phenomenon occurring. After literature retrieving, we found many of our resulting genes showed biological associations with this disease (see supplementary table 1). For example, when 790 genes were carried out functional enrichment studies on GO functions and pathway enrichments, these genes were prone to participate majorly in antigen processing and presentation, cell adhesion molecules (CAMs), glutathione metabolism and cytokine-cytokine receptor interaction and so on, even if dysfunction was found in any of these biological processes, it seems to have little effect on other processes and could not lead to systemically abnormality or impairment in the function of human essential immune system. And meanwhile, 98 genes out of our proposed method were too minor to be detected using GWAS and showed a significant and consistent state with known disease genes in function levels (see “Functional Annotation Screening and Candidate Gene Prediction” in method section). We propose that multi-dimensional approaches can also be applied to candidate gene identification of rheumatoid arthritis, where multiple genes sharing underlying biological similarities (e.g. in the same pathway or GO term) contribute to disease etiology but where common variations in each of these genes make modest contributions to disease risk. Considering underlying biological similarities together, rather than focusing on a few SNPs and/or genes with
the strongest evidence of disease association, we can detect likely-causing genes using the proposed method. According to the reviewer’s advice, we have incorporated the following elaborations in the revised paragraphs:

In the section of supplemental file:
“Supplementary figure 1 - the GO functional enrichment of candidate genes predicted by our proposed method.tif”
“Supplementary figure 2 - the GO functional enrichment of candidate genes out of GWAS.tif”
“Supplementary figure 3 - the GO functional enrichment of known disease genes.tif”
“Supplementary table 1 - candidate genes with their corresponding risk SNPs predicted by our method.xls”
“Supplementary table 4 - the KEGG functional enrichment of candidate genes predicted by our method.xls”
“Supplementary table 5 - the KEGG functional enrichment of known disease genes.xls”
“Supplementary table 6 - the KEGG functional enrichment of candidate genes out of GWAS.xls”

(7). “In pages 13 and 14, is it correct that "c" in equation #1 is the same as "c" in equation #4? The first one is for single binomial distribution, and the second one is for three independent binomial distributions? Do they have to be same?”

REPLY: Thank the reviewer to point out the insufficient description of the method section. In fact, this Bayesian association approach is to test for association with case-control phenotype at a single SNP locus. For a given SNP locus, we assume that data, D, are counts of cases and controls for each of the three genotypes at this locus, such as in table 2. We must denote that the matrix, D, is given six constant values (\( n_{0}^{A}, n_{0}^{v}, n_{1}^{A}, n_{1}^{v}, n_{2}^{A}, n_{2}^{v} \)) for a specific SNP beforehand. Under \( H_{0} \) the probability of the observed dataset, D (\( \{ n_{0}^{A}, n_{0}^{v}, n_{1}^{A}, n_{1}^{v}, n_{2}^{A}, n_{2}^{v} \} \)), does not depend on genotype, and can be written in terms of the probability \( \theta \) that an individual included in the study is a case:

\[
P(D / \theta) = C_{n}^{n_{0}^{A}} \theta^{n_{0}^{A}} (1 - \theta)^{n_{0}^{v}} \times C_{n}^{n_{1}^{A}} \theta^{n_{1}^{A}} (1 - \theta)^{n_{1}^{v}} \times C_{n}^{n_{2}^{A}} \theta^{n_{2}^{A}} (1 - \theta)^{n_{2}^{v}}
\]

\[
= C_{n}^{n_{0}^{A}} C_{n}^{n_{0}^{v}} C_{n}^{n_{1}^{A}} C_{n}^{n_{1}^{v}} C_{n}^{n_{2}^{A}} C_{n}^{n_{2}^{v}} \theta^{n_{0}^{A} + n_{1}^{A} + n_{2}^{A}} (1 - \theta)^{n_{0}^{v} + n_{1}^{v} + n_{2}^{v}}
\]

(1)

Table 2 Frequency of cases and controls for each of three genotypes at a SNP locus

<table>
<thead>
<tr>
<th>Genotype</th>
<th>AA</th>
<th>AB</th>
<th>BB</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>( n_{0}^{A} )</td>
<td>( n_{1}^{A} )</td>
<td>( n_{2}^{A} )</td>
<td>( n_{0}^{A} + n_{1}^{A} + n_{2}^{A} )</td>
</tr>
<tr>
<td>Control</td>
<td>( n_{0}^{v} )</td>
<td>( n_{1}^{v} )</td>
<td>( n_{2}^{v} )</td>
<td>( n_{0}^{v} + n_{1}^{v} + n_{2}^{v} )</td>
</tr>
<tr>
<td>Total</td>
<td>( n_{0}^{A} + n_{0}^{v} )</td>
<td>( n_{1}^{A} + n_{1}^{v} )</td>
<td>( n_{2}^{A} + n_{2}^{v} )</td>
<td>( n_{0}^{A} + n_{1}^{A} + n_{2}^{A} + n_{0}^{v} + n_{1}^{v} + n_{2}^{v} )</td>
</tr>
</tbody>
</table>

Where \( n_{0}^{A}, n_{0}^{v}, n_{1}^{A}, n_{1}^{v}, n_{2}^{A}, n_{2}^{v} \) represents the frequency of cases and controls for each of the specific genotype AA, AB or BB, respectively.

To compute a probability for the observed dataset, D under \( H_{1} \), we can assume that individuals with genotype \( j \) have probability \( \theta_{j} \) to be a case. Then analogous to (1) we can write

\[
P(D | \theta_{0}, \theta_{1}, \theta_{2}) = C_{n}^{n_{0}^{A}} \theta_{0}^{n_{0}^{A}} (1 - \theta_{0})^{n_{0}^{v}} \times C_{n}^{n_{1}^{A}} \theta_{1}^{n_{1}^{A}} (1 - \theta_{1})^{n_{1}^{v}} \times C_{n}^{n_{2}^{A}} \theta_{2}^{n_{2}^{A}} (1 - \theta_{2})^{n_{2}^{v}}
\]

\[
= C_{n}^{n_{0}^{A}} C_{n}^{n_{0}^{v}} C_{n}^{n_{1}^{A}} \theta_{0}^{n_{0}^{A}} (1 - \theta_{0})^{n_{0}^{v}} \times \theta_{1}^{n_{1}^{A}} (1 - \theta_{1})^{n_{1}^{v}} \times \theta_{2}^{n_{2}^{A}} (1 - \theta_{2})^{n_{2}^{v}}
\]

(3)

For a given SNP locus, a six fold table, D, is defined beforehand. We argue that "c" in equation #1 is the same as "c" in equation #4. Under a given matrix, D, the first one is for three genotypes with
the same probability \( \theta \), and the second one is for three genotypes with three individual probabilities \( (\theta_0, \theta_1, \theta_2) \). So they have to be the same and can be ignored during analysis.

(8). “In page 14, how do they guarantee that \( \theta_0 + \theta_1 + \theta_2 \) will be always 1 from equation #3?”

**REPLY:** Thank the reviewer to point it out. As the reviewer has pointed out, \( \theta_0 + \theta_1 + \theta_2 \) could not be guaranteed to be always 1 in equation #3. To take the easiest approach first, we do not impose any further constraints on \( \theta = (\theta_0, \theta_1, \theta_2) \), and thus our disease model is fully general, including for example dominance \( (\theta_0 = \theta_1 < \theta_2) \) and over-dominance \( (\theta_0, \theta_2 < \theta_1) \).

And furthermore, our primary interest is in the association of the SNP locus with the disease, and so we regard \( \theta_j (j = 0, 1, 2) \) as “nuisance” parameters, which, assuming uniform priors, can be eliminated by integrating equation #3 with respect to each \( \theta_j \) to give equation #4 as follows:

\[
P(D) = cB(n_0^A + 1, n_0^U + 1) \times B(n_1^A + 1, n_1^U + 1) \times B(n_2^A + 1, n_2^U + 1)
\]

In fact, the probability of D under \( H_1 \) is the approximation of its real value, \( \bar{P}(D)_{\text{real}} \). In this article, to reduce the computational complexity and make the final formula easier, we adopt an alternative metric, BF, to represent its real value \( \bar{BF}_{\text{real}} \). And the relationship between BF and \( \bar{BF}_{\text{real}} \) is illustrated as follows:

\[
\bar{P}(D)_{\text{real}} = \int \int \int P(D | \theta_0, \theta_1, \theta_2) d\theta_0 d\theta_1 d\theta_2 \\
\leq c \int \int \sum_{0 \leq \theta_0 \leq 1} \int \sum_{0 \leq \theta_1 \leq 1} \int \sum_{0 \leq \theta_2 \leq 1} \theta_0^{n_0^A} (1 - \theta_0)^{n_0^U} \theta_1^{n_1^A} (1 - \theta_1)^{n_1^U} \theta_2^{n_2^A} (1 - \theta_2)^{n_2^U} d\theta_0 d\theta_1 d\theta_2 \\
= cB(n_0^A + 1, n_0^U + 1) \times B(n_1^A + 1, n_1^U + 1) \times B(n_2^A + 1, n_2^U + 1) \\
= P(D)
\]

\[
BF = \frac{B(n_0^A + 1, n_0^U + 1) B(n_1^A + 1, n_1^U + 1) B(n_2^A + 1, n_2^U + 1)}{B(n_0^A + n_0^U + n_1^A + n_1^U + n_2^A + n_2^U + n_0^A + n_0^U + n_1^A + n_1^U + n_2^A + n_2^U + 1)} \\
\approx \frac{\bar{P}(D)_{\text{real}}}{P(D)_{\theta_0}} = \bar{BF}_{\text{real}}
\]

Here, we let \( BF > 1 \) to guarantee that all the risk SNPs could be included during our first genetic screening. And this approximate treatment could reduce the computational complexity. And we can regard \( \theta_j (j = 0, 1, 2) \) as “nuisance” parameters, without caring about the real value of the parameter of \( \theta_j (j = 0, 1, 2) \).

(9). “Minor problems: Before using abbreviations such as SVM, the authors need to define/describe them first.”

**REPLY:** Thank the reviewer to point it out. According to the reviewer’s advice, we have revised the corresponding description and rewritten the ‘SVM’ section in the revised version. We have incorporated the following elaborations in the revised paragraphs:

In the section of ‘Method section’:

**SVM Classification based on Sequence and Structure Similarity Features**

Firstly, ID Converter[48] is used to map all the genes to their corresponding proteins across the whole human genome. In this section, the positive set consists of disease genes of rheumatoid arthritis gathered from NCBI and OMIM online database (see in supplementary table 3). While, the negative set contains the remaining genes that won’t 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 (in table 3), a combinational pseudo-sequence, is used to represent each protein in positive, negative and testing sets according to the online RCSB PDB and targetDB databases. Here, 8-dimension secondary features (21-28) and the entire 28-dimension physicochemical features are used to train two classifiers and we carry out the second screening.

Considering the diversity of the putative non disease-candidate proteins, the non disease-candidate space might not have been sampled completely. Therefore, we construct 1,000 additional training sets (Positive/Negative=1:1), in which each negative set is selected randomly from the original negative set. During each randomization, the 8- and 28-dimension feature are used to construct the corresponding classifier. The performance of our model is evaluated with an n-fold cross-validation test. In the cross-validation test, the entire positive and negative data sets are shuffled and split into n folds. Each fold is used in turn for testing and the remaining part (n-1 folds) is used for training. The sensitivity ($Q_p$), specificity ($Q_n$) and overall accuracy ($Q_a$) are used to measure the accuracy of positive prediction, negative prediction and the overall accuracy of the model [49], respectively.

$$Q_p = \frac{TP}{TP + FN}$$

$$Q_n = \frac{TN}{TN + FP}$$

$$Q_a = \frac{TP + TN}{TP + TN + FP + FN}$$

Here, TP, TN, FP and FN represent true positives, true negatives, false positives and false negatives, respectively. In general, the overall accuracy $Q_a$ is always used to measure the predictive power of a model.

We construct 10,000 additional training sets (Positive/Negative=335:335), in which each negative set is selected randomly from the original negative set. Here, the test set is prepared from genetic screenings. We utilize the training sets, the test set and the optimal classifier to retrain the classifier. In this process, we aim to classify the genes in the test set, and then predict candidate genes of RA. We perform randomization 10,000 times and keep the genes which are judged to be disease genes in each process of random as our predictions.”

(10). “Minor problems: Figure 2 was not mentioned in the main text at all.”

**REPLY:** Thank the reviewer to point it out. According to the reviewer’s advice, we have referenced figure 2 at the right position in the revised version. We have incorporated the following elaborations in the revised paragraphs.

(11). “Minor problems: I believe "Bayes Factor" is the correct name, not "Bayesian Factor".”

**REPLY:** Thank the reviewer to point it out. According to the reviewer’s advice, we have revised the corresponding description of “Bayesian Factor” to be “Bayes Factor” in the corresponding section of the revised version. We have incorporated the following elaborations in the revised paragraphs:

In the section of ‘Methods’:
“...The next step is to compute the Bayes Factor (BF), which is the ratio of (4) to (2). The corresponding formula is as follows:
......”

(12). “Minor problems: In page 14, use consistent notation for the Beta function. Choose either B or Beta. Same thing for the numbers of cases and controls. Either "a" in equation #5 or "n" as in equation #3.”

**REPLY:** Thank the reviewer to point it out. According to the reviewer’s advice, we have revised the corresponding description and rewritten the ‘Methods’ section in the revised version. We have incorporated the following elaborations in the revised paragraphs:

In the section of ‘Methods’:
“formula 1-6” see in the revised manuscript.
Reply to Reviewer #2 (report from 25 February 2010, Reviewer: Nora L Nock)

(1) “Hypothesis & Conclusions: The specific research question posed by the authors is not well defined. The Abstract, in particular, does not adequately describe the authors’ approach, how this approach is novel and how the approach improves upon current findings. The authors state in the Abstract that “our method could return more ‘precise’ results”; however, no results comparing their method to a “traditional” genome-wide association analysis approach are provided in the Abstract. It appears the authors did perform a “traditional” genome-wide association (GWA) analysis of the Genetic Analysis Workshop (GAW) 16 Rheumatoid Arthritis (RA) data but the results of this are not mentioned in the Abstract (and the methodological details provided on p.17 are scant, see “Minor Revisions” below. Furthermore, although the authors discuss the “traditional” GWA results on p.10 and compare the methods using a Venn Diagram (Figure 5), the authors should also provide a list of the specific candidate genes (& specific variant rs# from which the genes was identified) for each method (Note: This information could be placed in Supplemental Materials if there is not room in the manuscript). From Figure 5, it appears only 22 candidate genes overlap between the two approaches; however, on p.10, the authors state their significance level on the “traditional” GWA was set to 0.01 (not the more widely accepted level of 1 X 10^-7 for a 500K panel). Therefore, the authors should discuss how changing the significance criteria would affect the results of this comparative analysis. Furthermore, the authors need to justify how they concluded that there are “34 known disease genes of RA” – is this based on individual candidate gene papers?”

REPLY: Thank the reviewer to pointing out the insufficient description of the ‘Abstract’ section and the method comparison section. And the specific candidate genes and specific variants rs# were listed in the supplementary table 1 and 2.

We accepted the reviewer’s suggestion and better explained and defined the specific research question. Our proposed approach to the analysis of GAW data is motivated by underlying biological similarities-based methods between candidate and known disease genes. When genetic profiles of different variances are compared, tens or hundreds of variances (SNPs) may have subtle differences in case-control phenotypes. Rather than focusing on individual genes that have the strongest evidence, our multi-dimensional screening approaches typically extract all risk genes (BFLn>0) by their odd ratios of hypothesis $H_1$ to $H_0$ and then look for whether a particular group of genes shares underlying biological similarities with known disease genes. Application of our method has yielded biological insights that are otherwise undetectable by focusing only on genes with the strongest evidence of multiple testing. A large part of candidate genes out of traditional prediction could not easily be classified into related function categories or interacting biological processes associated with this disease (supplementary figure 1, 2 and 3). For example, when 790 genes were carried out functional enrichment studies on GO functions and pathway enrichments, these
genes were prone to participate majorly in antigen processing and presentation, cell adhesion molecules (CAMs), glutathione metabolism and cytokine-cytokine receptor interaction and so on, even if dysfunction was found in any of these biological processes, it seems to have little effect on other processes and could not lead to systemically abnormality or impairment in the function of human essential immune system. And meanwhile, 98 genes out of our proposed method were too minor to be detected using GWAS and showed a significant and consistent state with known disease genes in function levels (see “Functional Annotation Screening and Candidate Gene Prediction” in method section). We propose that multi-dimensional approaches can also be applied to candidate gene identification of rheumatoid arthritis, where multiple genes sharing underlying biological similarities (e.g. in the same pathway or GO term) contribute to disease etiology but where common variations in each of these genes make modest contributions to disease risk. Considering underlying biological similarities together, rather than focusing on a few SNPs and/or genes with the strongest evidence of disease association, we can detect likely-causing genes using the proposed method. We have revised the corresponding section to make it readable.

The results comparing our method to a “traditional genome-wide association analysis” are provided in the supplementary figure 2, table 2 and table 6. Thank the reviewer very much for pointing out the error description in page 10. With respect to the description of significance level in GWAS, in fact the significance level we used on p.10 was set to 1.835e-8 (=0.01/S, S is the size of all the SNPs. this is the significance level of Bonferroni test). The value 0.01 we stated is the significant level of multiple testing.

The term of “34 known disease genes of RA” is also an error description. When the 146 genes were identified by our proposed method, we use the NCBI PubMed module to retrieve their association with this disease (rheumatoid arthritis) using the term “GENE symbols+rheumatoid arthritis” (e.g. CD4+rheumatoid arthritis). Finally, 58 genes are identified to be associated with this disease and more information could be found in supplementary table 1.

We have incorporated the following elaborations in the revised paragraphs:

In the section of “Abstract”:

“Abstract Background: Hundreds of thousands of SNPs have been tested for rheumatoid arthritis according to the Genetic Analysis Workshops (GAWs). Traditional genome-wide association studies (GWAS) have been developed to identify susceptibility genes assuming a “most significant SNPs/genes” model. And meanwhile, lots of minor or modest risk genes are likely to be missed after adjustment of multiple testing, and this screening process resorts to a strict selection of statistical thresholds, which aims to identify susceptibility genes based only on the statistic model without considering multi-dimensional biological similarities in sequence arrangement, crystal structures and functional categories/biological pathways between candidate and known disease genes. Multidimensional screening approaches, combined with traditional statistical genetics methods, could consider multiple biological backgrounds of genetic mutation, structural and functional annotations. Here we introduce a towards-multidimensional screening approach to predicate candidate genes of rheumatoid arthritis that take into account all SNPs with nominal evidence of Bayesian association (BFLn>0), structural and functional similarities of their corresponding genes or proteins.

Results: Traditional candidate gene prediction studies return 822 genes and a large part of these candidate genes could not easily be classified into related functional categories or biological processes associated to this disease, while our prediction could prevent this phenomenon occurring. Finally, we identified 146 candidate genes of rheumatoid arthritis including gene CD4, FGFR1 and KDR and so on, which were reported as high risk factors by most recent studies. And meanwhile, some minor or modest disease risk genes could be identified by our proposed method, and the associations between these genes and this disease are verified by literature retrieving.

Conclusions: Our proposed approach to the analysis of GAW16 data (rheumatoid arthritis) is motivated by underlying biological similarities-based methods between candidate and known disease genes. And it could identify candidate likely-causing disease genes of rheumatoid arthritis. We hope that using the proposed method could provide additional insights into the pathogenesis of rheumatoid arthritis and other diseases, when handling tens of hundreds of genetic variances datasets.”
In the section of supplemental file:

“Supplementary figure 1 - the GO functional enrichment of candidate genes predicted by our proposed method.tif”
“Supplementary figure 2 - the GO functional enrichment of candidate genes out of GWAS.tif”
“Supplementary figure 3 - the GO functional enrichment of known disease genes.tif”
“Supplementary table 1 - candidate genes with their corresponding risk SNPs predicted by our method.xls”
“Supplementary table 4 - the KEGG functional enrichment of candidate genes predicted by our method.xls”
“Supplementary table 5 - the KEGG functional enrichment of known disease genes.xls”
“Supplementary table 6 - the KEGG functional enrichment of candidate genes out of GWAS.xls”

(2). “Acknowledgement of Prior Work (Introduction & Discussion): The authors do not adequately address and reference prior methodological work. For example, on page 4, the authors’ state: “A limited number of studies have used genome-wide association studies, function clustering algorithms or pattern recognition methods based on structural genomics knowledgebase of disease and non-disease genes to identify candidate genes, respectively. However, these studies to date do not identify candidate genes from multidimensional genomic annotations or knowledgebase.” However, these statements are not substantiated with supporting details or references.

In addition, in the Results and Discussion section on p.6, the authors’ state: “Consistent with all the previous associated studies of rheumatoid arthritis, genes in the strongest gene set were members of the immunoglobulin protein family, protein kinase domain family, SH3 domain family and ligand-binding domain of nuclear hormone receptor family and so on, and they included several high disease risk and hotly-reported genes such as CD4, FGFR1 and KDR.”; however, they provide no supporting references.”

REPLY: Thank the reviewer to point it out. We have provided supporting references in the revised manuscript and address adequately and reference prior methodological work. We have incorporated the following elaborations in the revised paragraphs:

In the section of ‘Background’:

“A limited number of studies have used genome-wide association studies [23-25], function clustering algorithms [26-29] or machine learning methods based on structural genomics knowledgebase [30] to identify candidate genes of rheumatoid arthritis, respectively. However, when a set of candidate risk genes are acquired from case-control datasets of genetic variances, jointly considering structural and functional associations between candidate genes and disease to identify candidate genes might provide additional insights into the results of traditional statistic genetics analysis.”

In the section of “Results and Discussion section”:

“Due to the nature of our screening approach, many of our predictions overlap extensively in sorts of similar function categories. Thus, to describe function representative of association with rheumatoid arthritis, we selected those with the strongest association which also displayed a higher functional enrichment. For example, consistent with all the previous associated studies of rheumatoid arthritis, genes in our resulting gene set were members of the immunoglobulin protein family (see in figure 2), protein kinase domain family, SH3 domain family and ligand-binding domain of nuclear hormone receptor family and so on, and they included several moderate disease risk and hotly-reported genes such as CD4 [32-35], FGFR1 [36-38] and KDR [39-43]. These genes in the immunoglobulin protein family played a crucial role [44, 45] in the pathogenesis of this disease:

… …”
(3). “Databases: The authors need to provide the accession dates and versions/builds for the various databases utilized since the information in these databases is continually being updated and could affect their results and conclusions.”

**REPLY:** Thank the reviewer very much for pointing out the insufficient description about the accession dates or versions for the various databases. We have incorporated the following elaborations in the revised paragraphs:

In the section of ‘Materials and Methods’:

- **Gene Location and Disease Loci Data:** Location information of the human genes was acquired from the NCBI genome database (downloaded on Mar 25, 2009). Disease loci information was gathered by OMIM online database (downloaded on Mar 25, 2009) [46].
- **Sequence and Crystal Structure Data:** The linear-sequence items of all the human genes are obtained from NCBI genome database. And the crystal structure datasets of human proteomics are collected from online databases, called 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 are used as source function annotations. And these well-defined categories are widely introduced to important function identification analysis. In this study, each candidate gene is annotated onto its corresponding function families or categories according to these three databases.

(4). “Minor problems: General: There are various typos and spacing issues through the document – too numerous to list here. The authors should also define ALL acronyms at first use. The transitions between sentences are very weak making it difficult to determine critical points the authors are trying to convey. The authors should carefully proof their manuscript and revise accordingly.

**REPLY:** Thank the reviewer very much for pointing out the minor problems about various typos and spacing issues through the document. According to the reviewer’s advice, we have carefully proofed our manuscript and revised weak descriptions.

(5). “Minor problems: Abstract: The Abstract is, perhaps, the most important part of the manuscript. The authors should revise the Abstract to make it clearer and more informative. For example, the authors state that “modest risk genes are likely to be missed after adjustment for multiple testing” (presumably using “traditional” GWA analysis approaches); however, the authors never provide specific results comparing the standard “GWA” analyses approaches to their method nor do they clarify how their method improves upon the multiple testing issue.

**REPLY:** Thank the reviewer very much for pointing out the insufficient description of the ‘Abstract’ section. Our proposed approach to the analysis of GAW data is motivated by underlying biological similarities-based methods between candidate and known disease genes. When genetic profiles of different variances are compared, tens or hundreds of variances (SNPs) may have subtle differences in case-control phenotypes. Rather than focusing on individual genes that have the strongest evidence, our multi-dimensional screening approaches typically extract all risk genes (BFLn>0) by their odd ratios of hypothesis $H_1$ to $H_0$ and then look for whether a
particular group of genes share underlying biological similarities with known disease genes. Application of our method has yielded biological insights that are otherwise undetectable by focusing only on genes with the strongest evidence of multiple testing. A large part of candidate genes out of traditional prediction could not easily be classified into related function categories or interacting biological processes associated with this disease (supplemental figure 2 and table 6). For example, when 790 genes were carried out functional enrichment studies on GO functions and pathway enrichments, these genes were prone to participate majorly in antigen processing and presentation, cell adhesion molecules (CAMs), glutathione metabolism and cytokine-cytokine receptor interaction and so on, even if dysfunction was found in any of these biological processes, it seems to have little effect on other processes and could not lead to systemically abnormality or impairment in the function of human essential immune system. And meanwhile, 98 genes out of our proposed method were too minor to be detected using GWAS and showed a significant and consistent state with known disease genes in function levels (see “Functional Annotation Screening and Candidate Gene Prediction” in method section). We propose that multi-dimensional approaches can also be applied to candidate gene identification of rheumatoid arthritis, where multiple genes sharing underlying biological similarities (e.g. in the same pathway or GO term) contribute to disease etiology but where common variations in each of these genes make modest contributions to disease risk. Considering underlying biological similarities together, rather than focusing on a few SNPs and/or genes with the strongest evidence of disease association, we can detect likely-causing genes using the proposed method. According to the reviewer’s advice, we have incorporated the following elaborations in the revised paragraphs:

In the section of ‘Abstract’:

“Abstract Background: Hundreds of thousands of SNPs have been tested for rheumatoid arthritis according to the Genetic Analysis Workshops (GAWs). Traditional genome-wide association studies (GWAS) have been developed to identify susceptibility genes assuming a “most significant SNPs/genes” model. And meanwhile, lots of minor or modest risk genes are likely to be missed after adjustment of multiple testing, and this screening process resorts to a strict selection of statistical thresholds, which aims to identify susceptibility genes based only on the statistic model without considering multi-dimensional biological similarities in sequence arrangement, crystal structures and functional categories/biological pathways between candidate and known disease genes. Multidimensional screening approaches, combined with traditional statistical genetics methods, could consider multiple biological backgrounds of genetic mutation, structural and functional annotations. Here we introduce a towards-multidimensional screening approach to predicate candidate genes of rheumatoid arthritis that take into account all SNPs with nominal evidence of Bayesian association (BFLn>0), structural and functional similarities of their corresponding genes or proteins.

Results: Traditional candidate gene prediction studies return 822 genes and a large part of these candidate genes could not easily be classified into related functional categories or biological processes associated to this disease, while our prediction could prevent this phenomenon occurring. Finally, we identified 146 candidate genes of rheumatoid arthritis including gene CD4, FGFR1 and KDR and so on, which were reported as high risk factors by most recent studies. And meanwhile, some minor or modest disease risk genes could be identified by our proposed method, and the associations between these genes and this disease are verified by literature retrieving.

Conclusions: Our proposed approach to the analysis of GAW16 data (rheumatoid arthritis) is motivated by underlying biological similarities-based methods between candidate and known disease genes. And it could identify candidate likely-causing disease genes of rheumatoid arthritis. We hope that using the proposed method could provide additional insights into the pathogenesis of rheumatoid arthritis and other diseases, when handling tens of hundreds of genetic variances datasets.”

(6). “Minor problems: Conclusion: It appears the authors’ method could be utilized in other diseases; however, the authors never comment about this. It would be helpful for the authors to discuss this, if true, in the Discussion section (and Abstract).

REPLY: Thank the reviewer very much for pointing it out. Our proposed approach to the
analysis of GAW16 data (rheumatoid arthritis) is motivated by underlying biological similarities-based methods between candidate and known disease genes. In this article, we try to introduce a multi-dimensional screening approach analyze the GAW16 data and aim to identify responsible genes of rheumatoid arthritis as the manuscript tile mentioned. And moreover, we hope that the proposed method could provide additional insight into the pathogenesis of other diseases, when handling tens or hundreds of genetic variances datasets.

In the section of “Abstract”:

“Our proposed approach to the analysis of GAW16 data (rheumatoid arthritis) is motivated by underlying biological similarities-based methods between candidate and known disease genes. And it could identify candidate likely-causing disease genes of rheumatoid arthritis. We hope that using the proposed method could provide additional insights into the pathogenesis of rheumatoid arthritis and other diseases, when handling tens of hundreds of genetic variances datasets.”

In the section of “Results and Discussion”:

Multi-dimensional approaches can also be applied to candidate gene identification of other diseases, where multiple genes sharing underlying biological similarities (e.g. in the same pathway or GO term) contribute to disease etiology but where common variations in each of these genes make modest contributions to disease risk. Considering underlying biological similarities together, rather than focusing on a few SNPs and/or genes with the strongest evidence of disease association, researches can detect likely-causing genes using the proposed method. We hope that the proposed method could provide additional insights into the pathogenesis of other diseases, when handling tens of hundreds of genetic variances datasets.”

(7). “Minor problems: Methods: On p. 13, Table 1 is not “below”; and, it would be helpful to reference the reader to Table 1 again on p.14 where the variables enter into equations #2 and #5 (since the variables are not defined in the text). On p.17, it is unclear if the carrot (“^”) refers to an “and” or an “or” command. That is, does the gene need to be identified in just one or all three of the databases? It appears from Figure 1 that the “^” indeed refers to an “and” command; but, in the text on p.17, it says that “…we use f(gi)...by retaining genes that share one or more similar function annotations”. Thus, this issue needs clarified. Furthermore, the authors should comment on how the results might change if the algorithm was modified to include genes in only 1 (“or”), 2 or all 3 (“and”) databases. On p.17, additional details regarding the “traditional” GWA should be provided.

REPLY: Thank the reviewer very much for pointing out the insufficient description. In the section of “Functional Annotation Screening and Candidate Gene Prediction”, we hypothesized that disease genes would gather in some specific protein families, participate in the same biological functions or interact within some biological pathways. Thank the reviewer very much for pointing out the inconsistent and error description about the symbol ‘\&’. In fact, candidate genes identified by the algorithm include genes in only 1 (“or”), 2 or all 3 (“and”) databases. And we have revised the symbol ‘\&’ to ‘\&’. Additional details regarding the “traditional GWA” have been provided in supplementary figure 2, table 2 and table 6.

In the section of “Functional Annotation Screening and Candidate Gene Prediction”:
Finally, we use function $f(g_i)$ to carry out our functional annotation screening by retaining genes ($f(g_i) = 1$) that share at least one similar function annotation.

<table>
<thead>
<tr>
<th></th>
<th>$f_{PRISF}$</th>
<th>$f_{GO}$</th>
<th>$f_{KEGG}$</th>
<th>$f = f_{PRISF} \lor f_{GO} \lor f_{KEGG}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_1$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$g_2$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$g_3$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>$g_n$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 4 Sample screening. Rows represent the resulting genes of SVM screening and columns are function annotation information from PRISF, GO and KEGG databases. ‘$\lor$’ is an ‘or’ command.

In the section of “Comparison with Traditional Genome-wide Association Studies”

“Comparison with Traditional Genome-wide Association Studies

The traditional GWAs analysis[54] is carried out using the fisher exact test and multiple testing adjustment. Functional enrichments in GO biological processes and KEGG pathways are carried out for known disease genes, GWAs genes and our predicted genes, respectively. Functional consistence with known disease genes is examined to evaluate GWAs genes and our predicted genes. And to further evaluate the performance of our screening method, we utilize the NCBI PubMed module to retrieve the associations of GWAS genes or our resulting genes with this disease (rheumatoid arthritis) using the term “GENE symbols+rheumatoid arthritis” (e.g. CD4+rheumatoid arthritis) to depict the underlying mechanisms of our resulting genes.”

(8). “Minor problems: Figures: The figures need renumbered to avoid confusion for the reader (e.g., Figure 2 is on page 7 but Figure 1 is on p. 17). Figure 3A and 3B should be linked together.

REPLY: Thank the reviewer very much for pointing out it. And we have renumbered the figures and linked figure3A and 3B together.