

Research Article

A Meta-Analysis Strategy for Gene Prioritization Using Gene Expression, SNP Genotype, and eQTL Data

Jingmin Che and Miyoung Shin

Bio-Intelligence & Data Mining Lab, School of Electronics Engineering, Kyungpook National University, 1370 Sankyuk-dong, Buk-gu, Daegu 702-701, Republic of Korea

Correspondence should be addressed to Miyoung Shin; shinmy@knu.ac.kr

Received 26 August 2014; Revised 20 October 2014; Accepted 21 October 2014

Academic Editor: Mingyue Zheng

Copyright © 2015 J. Che and M. Shin. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to understand disease pathogenesis, improve medical diagnosis, or discover effective drug targets, it is important to identify significant genes deeply involved in human disease. For this purpose, many earlier approaches attempted to prioritize candidate genes using gene expression profiles or SNP genotype data, but they often suffer from producing many false-positive results. To address this issue, in this paper, we propose a meta-analysis strategy for gene prioritization that employs three different genetic resources—gene expression data, single nucleotide polymorphism (SNP) genotype data, and expression quantitative trait loci (eQTL) data—in an integrative manner. For integration, we utilized an improved technique for the order of preference by similarity to ideal solution (TOPSIS) to combine scores from distinct resources. This method was evaluated on two publicly available datasets regarding prostate cancer and lung cancer to identify disease-related genes. Consequently, our proposed strategy for gene prioritization showed its superiority to conventional methods in discovering significant disease-related genes with several types of genetic resources, while making good use of potential complementarities among available resources.

1. Introduction

The recent advance in high-throughput experiment technologies like microarrays and next-generation sequencing technologies has led to the production of large amounts of various biological resources regarding human genetic and disease-oriented data. Thus, it became one of the most significant issues in current biomedical research to identify disease genetic markers by exploring such a variety of resources in a systematic way. For this purpose, many earlier works [1–10] have been done by prioritizing candidate genes based on gene expression profiles or SNP genotype data, but they often produce many false-positive results, leading to the increase of time and cost to validate them experimentally.

In differential gene expression studies, the most common approach for gene prioritization is to utilize statistical methods for case-control microarray data which include t -test and significance analysis of microarrays (SAM) [1]. In these methods, candidate genes are prioritized according

to P values and disease markers are chosen as such genes that have P values lower than a specific threshold. Other methods like fold change or information gain are also used to select probable disease-associated genes. On the other hand, genome-wide association studies (GWAS) are often made to identify genetic variations associated with specific diseases. For this purpose, some statistical methods, like the Cochran-Armitage trend test (CATT) [2, 3], the genotypic χ^2 test, and the allelic χ^2 test [4], are widely used. The CATT based on a specific genetic model usually performs better than Pearson's χ^2 test with 2 degrees of freedom [5] and has therefore been suggested for use in the analysis of case-control data [6]. Since the underlying genetic models of a complex disease are often unknown, the CATT is widely used in combination with an additive model. Although GWAS can identify SNPs and other variations in DNA that are associated with specific diseases, they cannot determine specific causal genes [7, 8]. In order to link the SNP-level data to the gene-level data, Lehne et al. [9]

proposed the use of MaxT, MeanT, and TopQ since each gene might have several SNPs assigned to it.

There have also been extensive studies to examine expression quantitative trait loci (eQTL), which regulate mRNA and protein expression levels [10]. The eQTL can provide great insights into the molecular mechanisms underlying complex traits and aid in elucidating regulatory networks [11]. Furthermore, because eQTL data allow for the mapping of SNPs to biologically relevant genes [12], the Sherlock algorithm [13] employed the SNP and eQTL data to discover potential disease genes.

In spite of many positive aspects, however, these methods for gene prioritization have some disadvantages. For example, differential gene expression studies and GWAS focus only on a single data type for gene prioritization, so they suffer from being limited to a single genetic resource. This results that some potential disease-associated genes identified in differential gene expression studies remain undetected in GWAS, and vice versa. Also, both of them tend to show high false-positive rates. Thus, lately, an increasing number of gene prioritization tools are interested in integrating data from several resources [14–16].

In this paper, we propose a meta-analysis strategy for gene prioritization which integrates three different genetic resources, namely, gene expression data, SNP genotype data, and eQTL data, with an improved technique for order of preference by similarity to ideal solution (TOPSIS) [17]. The key idea of the proposed strategy is to utilize additional gene-level data obtained by using the eQTL data that provides SNP-gene mapping relationship and to combine the significance scores of candidate genes from three genetic resources with the improved TOPSIS. Our experiment results showed excellent performance of the proposed strategy in discovering significant disease-related genes.

2. Materials and Methods

2.1. Test Methods for Individual Genetic Resources. For the significance testing of gene expression data and SNP genotype data, we used the t -test and the CATT, respectively, which are the most commonly used ones in differential gene expression studies and GWAS, respectively.

2.2. Methods for Filtering Out Duplicate Gene Scores

Max. It is to choose the best score from duplicate gene scores for a certain gene. For example, in the case of t -test, SAM, and CATT analyses, the smallest P value is selected, and for information gain, the largest score is selected.

TopQ. It is to select the best quartile of all the duplicate scores given to a gene and use their arithmetic mean as a representative score of the gene. Here, if the number of the duplicate scores given to a gene is not a multiple of 4, the quartile number should be rounded up to the next integer.

Mean. It is to calculate the arithmetic mean of duplicate gene scores as a representative score of a certain gene.

2.3. Methods for Integrating Scores from Different Genetic Resources

The Improved TOPSIS Method. The original TOPSIS [18] is a method to measure comprehensive benefit of an object based on its relative distance to the ideal solution. The basic idea of this method is to find the positive ideal solution and the negative ideal solution in a decision-making process and then choose the alternatives in the descending order of the similarities to the positive ideal solution and in the ascending order of the distance to the negative ideal solution. On the other hand, the improved TOPSIS, which is a modified version of the original TOPSIS, was proposed as an evaluation method for economic problem [17]. To adapt this method for gene prioritization problem, we slightly modified the improved TOPSIS method. The detailed description of the procedure is given in the following.

- (1) Firstly, construct a $n \times m$ gene score matrix where n is the total number of genes and m is the number of gene scores obtained from different genetic resources:

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1m} \\ X_{21} & X_{22} & \cdots & X_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nm} \end{bmatrix}. \quad (1)$$

There can be some values missing in the matrix when the corresponding genes might not exist in the score list of some of the genetic resources. However, the improved TOPSIS method is not affected by missing values because it simply integrates existing scores only for specific genes.

- (2) Secondly, normalize the gene scores in each individual resource by dividing each of them by the Euclidean norm of all the gene scores from the same resource, as in formula (2); for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$,

$$U_{ij} = \frac{X_{ij}}{\sqrt{\sum_{i=1}^n X_{ij}^2}}. \quad (2)$$

- (3) Thirdly, obtain the *most* positive solution (U_j^+) and the *most* negative solution (U_j^-) for each type of genetic resources. That is, for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$, calculate the following:

$$\begin{aligned} U^+ &= (U_1^+, U_2^+, \dots, U_m^+), & U_j^+ &= \max \{U_{ij}\}, \\ U^- &= (U_1^-, U_2^-, \dots, U_m^-), & U_j^- &= \min \{U_{ij}\}. \end{aligned} \quad (3)$$

Here the max indicates the selection of the most positive solution which is the best score chosen by taking the smallest P value from the t -test, SAM, or CATT results or taking the largest score from the results of information gain. On the other hand, the min indicates the selection of the most negative

solution which is the worst score chosen by taking the largest P value for the t -test, SAM, or CATT results or taking the smallest score for the results of information gain.

- (4) Finally, for each gene $i = 1, 2, \dots, n$, calculate its relative distance d_i to the most negative solution by using formula (4), and then select the genes which have the larger values of d_i to find more significant genes:

$$d_i = \frac{\langle \Delta U_i, \Delta U \rangle}{\|\Delta U\|^2}, \tag{4}$$

where \langle, \rangle indicates the inter product and $\|\cdot\|$ is the Euclidean norm. Consider

$$\begin{aligned} \Delta U_i &= (U_i - U^-), \\ \Delta U &= (U^+ - U^-), \\ \|\Delta U\| &= \sqrt{\sum_{j=1}^m (U_j^+ - U_j^-)^2}. \end{aligned} \tag{5}$$

Rank Product (see [19]). This method is to combine ranked lists for prioritization by using the following formula:

$$RP_g = \left(\prod_i^k r_{g,i} \right)^{1/k}, \tag{6}$$

where $r_{g,i}$ is the rank of gene g in the score list of the i th genetic resource. That is, for each gene, it computes the rank product via the geometric mean of the ranks in the score lists of different genetic resources. Then the rank product is used as a final score for gene prioritization.

Fisher’s Method (see [20]). This method combines extreme value probabilities from several tests, commonly known as P values, into one test statistic (χ^2) using the formula given in the following:

$$\chi_{2k}^2 \sim -2 \sum_{i=1}^k \ln(p_i). \tag{7}$$

Rescaled Sum of Z-Scores (see [21]). This method combines several individual Z-scores by using the formula of

$$RSZ = \frac{\sum_{i=1}^k Z_i}{\sqrt{k}}, \tag{8}$$

where k is the number of Z-scores to be combined.

3. Results and Discussion

3.1. Evaluation of Gene Prioritization Results Obtained by Integrating Genetic Resources with Improved TOPSIS. To evaluate

the proposed strategy of gene prioritization integrating different genetic resources with the improved TOPSIS, we made experiments with two different datasets regarding prostate cancer and lung cancer. For each dataset, we downloaded gene expression data, SNP genotype data, and eQTL data from publicly available databases. In particular, for prostate cancer data, we used the gene expression profiles of the GSE6919 dataset [22] which includes 128 samples of 65 cases and 63 controls as in [23]. For SNP genotype data, we used the GSE18333 dataset [24] excluding 10 United Kingdom samples from original 82 samples, and this left us with 72 Chinese samples of 39 cases and 33 controls. For lung cancer data, we used the GSE19804 dataset [25], which includes 120 samples of 60 cases and 60 controls, as the gene expression data. For SNP genotype data, we used the GSE33355 dataset [26] of 122 samples with 61 cases and 61 controls. Also, Affymetrix 6.0 eQTL data were used which are downloadable from SCANDb [27]. Finally, for the validation of gene prioritization results, we downloaded the details of disease-associated genes from the Gene Association Database (GAD) and found 786 prostate cancer related genes and 731 lung cancer related genes, respectively.

The overall procedure of the proposed strategy for gene prioritization is illustrated in Figure 1. To begin with, we preprocessed the gene expression data for specific disease by using the comprehensive robust multiarray average [28] method and produced the prostate cancer gene expression data consisting of 12,625 probes and 128 samples with 65 cases and 63 controls, and the lung cancer gene expression data consisted of 54,675 probes and 120 samples with 60 cases and 60 controls. For the processing of SNP genotype data, we removed such SNPs satisfying minimum allele frequency <0.01 and Hardy-Weinberg equilibrium test statistic value lower than ~ 7 . Consequently, we obtained the prostate cancer SNP genotype data consisting of 709,216 SNPs and 72 samples with 39 cases and 33 controls, and the lung cancer SNP genotype data consisted of 760,716 SNPs and 122 samples with 61 cases and 61 controls. Next, with the above preprocessed data of gene expression and SNP genotype, we converted the probe IDs (or SNP IDs) to gene symbols with gene (or SNP) annotations, producing two datasets named GeneExp data and GeneSNP data, respectively. Also, by using eQTL data that conveys the biological relationships between SNPs and their regulated genes, we converted SNP IDs in the SNP genotype data to gene symbols, producing another dataset named GeneQTL data. Thus, eventually, it resulted in generating three datasets of genes (i.e., GeneExp data, GeneSNP data, and GeneQTL data), where each dataset may contain duplicate genes occurring by multiple probes mapped into the same gene symbol. These duplicate genes, if any, were filtered out after obtaining gene scores for each dataset.

In order to obtain gene scores, we applied the most common test methods for the three datasets, respectively. The t -test was used for GeneExp data and the CATT was used for GeneSNP and GeneQTL data. Instead of these methods, any other common methods can be applicable for each dataset. Now, we have three datasets of gene scores from the GeneExp, GeneSNP, and GeneQTL data, which are named Gene scores, SNP scores, and eQTL scores, respectively.

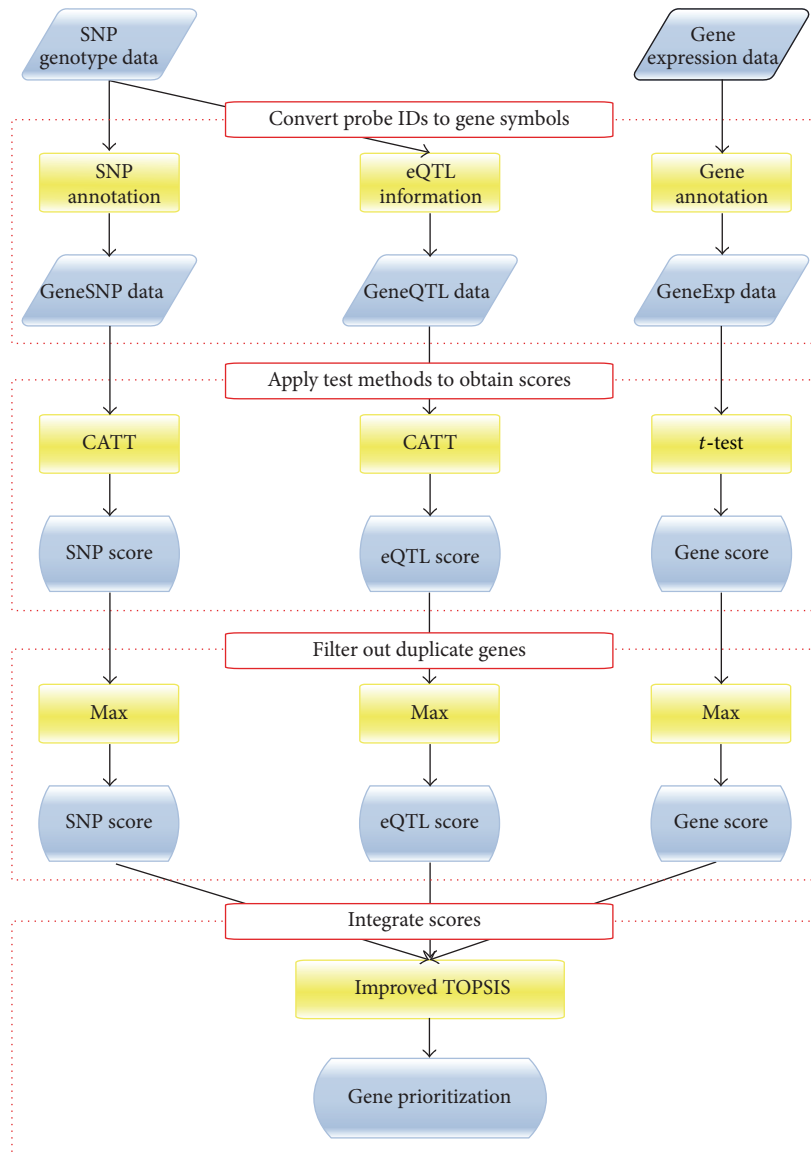


FIGURE 1: Overall procedure of the proposed strategy for gene prioritization which consists of four steps: (1) convert probe IDs to gene symbols, (2) apply test methods to obtain scores, (3) filter out duplicate genes in each score list, and (4) integrate scores with improved TOPSIS.

Based on these, we filtered out the duplicate genes by using one of the three available methods [9]: Max, TopQ, and Mean. In our experiments, we applied all three methods to remove duplicate genes and compared them in terms of the ability to discover potential disease-associated genes. Specifically, we chose the top 10% genes from each score list (ranked by P values) and, among them, counted the number of *actual* disease-related genes. The results from prostate cancer data and lung cancer data are shown in Figures 2(a) and 2(b), respectively, which clearly demonstrate that the Max method is the most suitable for filtering out duplicate genes. Thus, the Max method was used in all subsequent analyses. Consequently, after filtering out duplicate genes, we could obtain the prostate cancer dataset that includes 9,072 gene scores in the GeneExp data, 21,243 gene scores in the GeneSNP data, and 11,860 gene scores in the GeneQTL data.

Similarly, for the lung cancer dataset, we obtained 22,635 gene scores in the GeneExp data, 21,393 gene scores in the GeneSNP data, and 11,860 genes scores in the GeneQTL data.

Finally, with these three kinds of gene scores, we applied the improved TOPSIS method to integrate them and prioritized candidate genes according to the combined score. It should be noted that the candidate genes here to be prioritized are as many as the union of the genes in GeneExp data, GeneSNP data, and GeneQTL data, which leads to the maximal use of distinct genetic resources.

Our experiment results of gene prioritization are summarized in Figures 2, 3, and 4, where the performance was evaluated in terms of the receiver operating characteristic (ROC) curves and the area under the curve (AUC) estimates. In particular, Figure 3 shows the effects of integrating distinct genetic resources with improved TOPSIS on disease-related

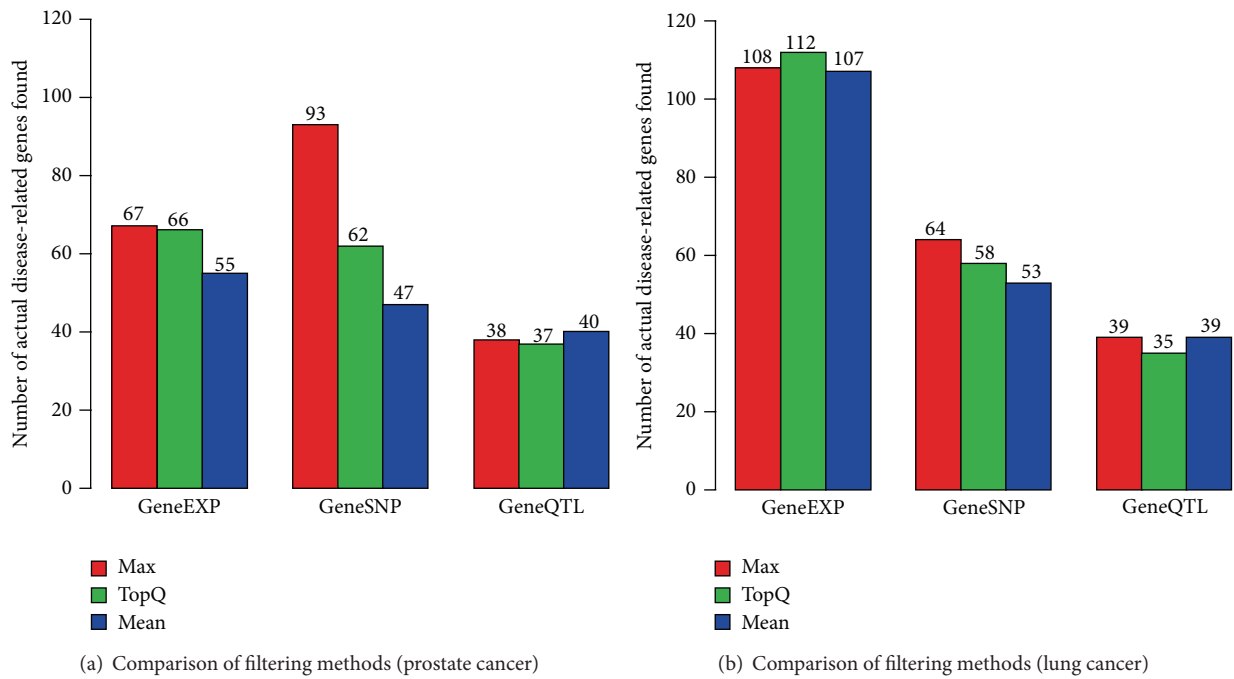


FIGURE 2: Comparison of the Max, TopQ, and Mean methods in filtering out duplicate genes: (a) prostate cancer results, (b) lung cancer results.

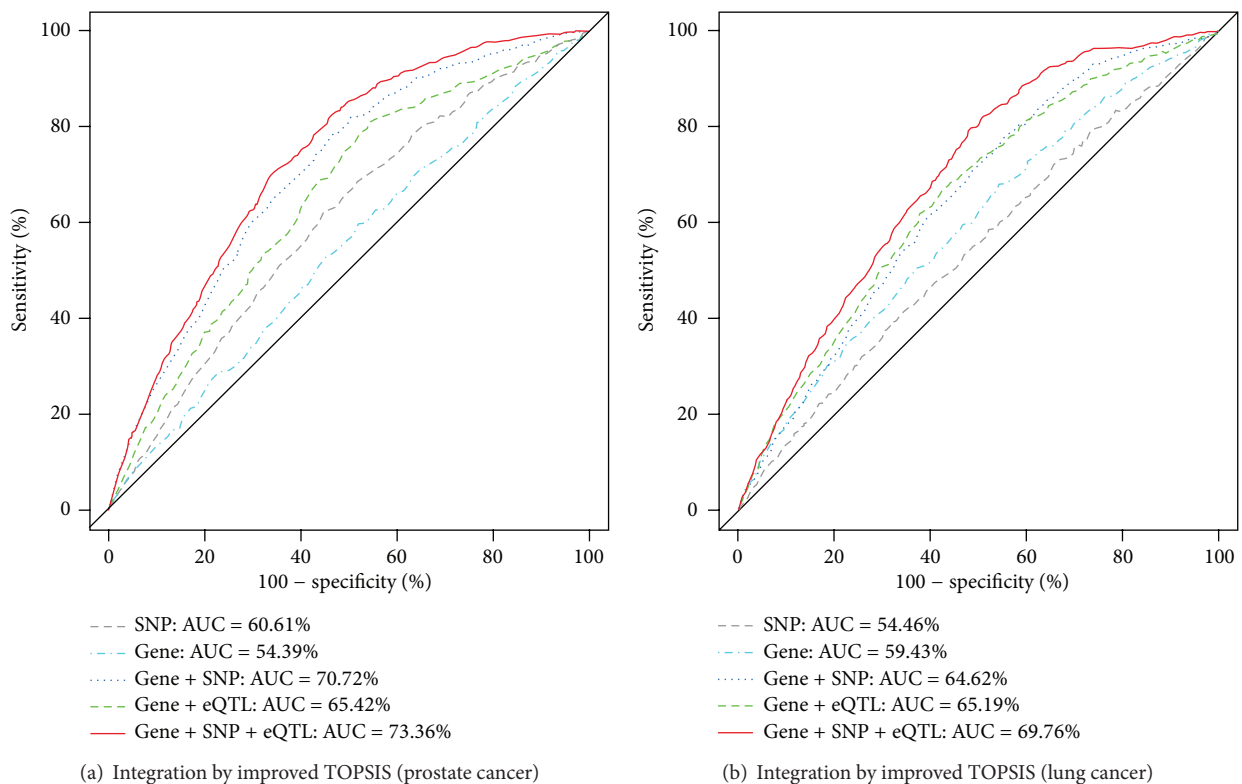


FIGURE 3: Effects of integrating distinct genetic resources with improved TOPSIS on disease-related gene identification: (a) prostate cancer results, (b) lung cancer results.

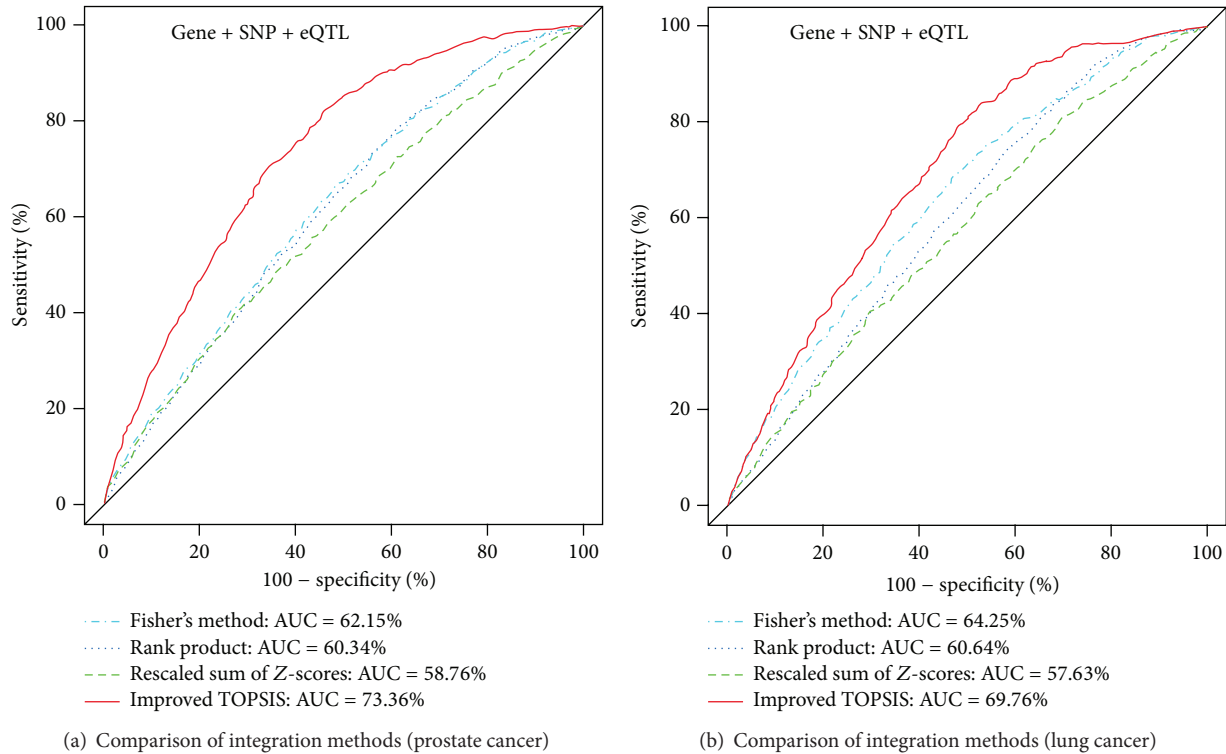


FIGURE 4: Comparison of the improved TOPSIS with other integration methods in terms of the ability to discover actual disease-related genes: (a) prostate cancer results, (b) lung cancer results.

gene identification in (a) prostate cancer data and (b) lung cancer data, respectively. From these figures, it is observed that the increasing number of distinct genetic resources to be used can be quite helpful to improve the performance of discovering potential disease-related genes, especially when the improved TOPSIS is used for the integration of different resources. In addition, this method does not only have the ability to cover as many genes as the union of the genes in different resources, but also can make good use of the potential complementarities among them.

3.2. Comparison of the Improved TOPSIS with Other Integration Methods. For the evaluation of our integrative approach with the improved TOPSIS, we also tested other integration methods (i.e., rank product, Fisher's method, and rescaled sum of Z-scores) under the same environment as in our experiments with the improved TOPSIS. Figure 4 shows the comparison of the improved TOPSIS with other integration methods in terms of the ability to discover actual disease-related genes in (a) prostate cancer data and (b) lung cancer data. According to these results, the improved TOPSIS performed much better in integrating scores from three distinct genetic resources (Gene, SNP, eQTL data) than the other methods. This may be the reason that only the improved TOPSIS can provide higher ranks to the genes found in all the three genetic resources than those found in a single resource or any two resources, whereas the other methods cannot do so. From the formulas of the rank product, Fisher's method, and rescaled sum of Z-scores, which are introduced

in the Methods, we can understand how such results can be obtained. For example, consider the case of two genes, in which the first gene's rank list is (1, 2, 3) and second gene's rank list is (3, NA, 1), where "NA" means that the gene is not present in the second genetic resource. When applying the rank product method to this case, the first gene's rank product is 1.82 and the second gene's rank product is 1.73. As a result, this method places the second gene in higher rank than the first gene, even though the first gene is actually much more important because it is present in all genetic resources. The Fisher and rescaled sum of Z-scores methods have similar problems. Consequently, it seems that such integration methods like the rank product method, Fisher's methods, and rescaled sum of Z-scores, are not suitable for integrating scores from these types of genetic resources.

3.3. Comparison of Our Strategy with Other Gene Prioritization Tools. For comparative purpose, we performed similar experiments with two existing meta-analysis tools for gene prioritization, MetaRanker 2.0 and Sherlock. The MetaRanker 2.0 is a web-based gene prioritization tool in which several types of data from different genetic resources can be given as inputs. For our analyses with this tool, we used the same three genetic resources as in our earlier experiments, including SNP genotype data, gene expression data, and eQTL data. On the other hand, the Sherlock is a tool to discover disease-related genes via genome-wide association study using eQTL information. For experiments with the Sherlock, we used SNP genotype data (which is the same as

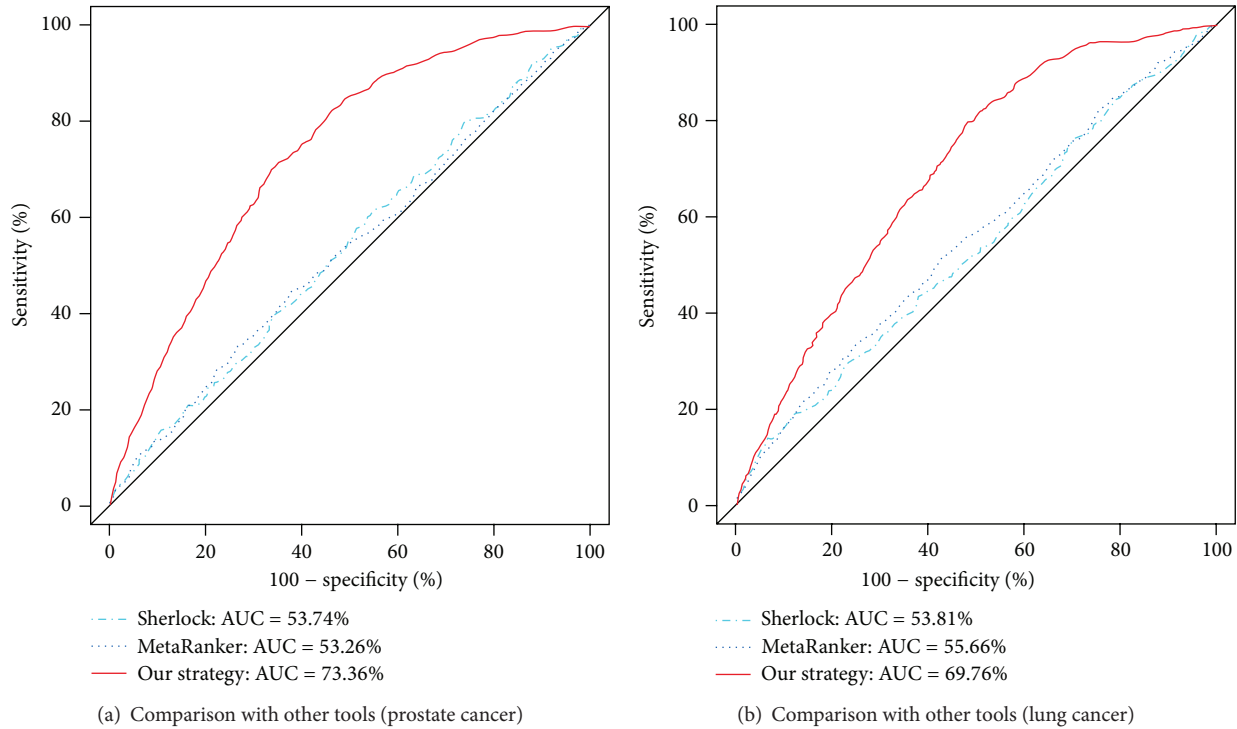


FIGURE 5: Comparison of our gene prioritization strategy with other meta-analysis tools in identifying actual disease-related genes: (a) prostate cancer results, (b) lung cancer results.

in our earlier experiments) and the eQTL data used in [29] which is available to choose at the webpage of the Sherlock. Figure 5 shows the comparison of our proposed strategy with these tools in identifying disease-related genes from (a) prostate cancer data and (b) lung cancer data. From these figures, it can be clearly observed that our integrative strategy for gene prioritization is superior to other meta-analysis tools, such as Sherlock and MetaRanker 2.0. Specifically, for prostate cancer data, our strategy showed 73.36% AUC estimate in identifying disease-related genes while the Sherlock and the MetaRanker 2.0 showed 53.74% and 53.26% AUC estimates respectively. Similarly, our strategy showed 69.76% AUC estimate in lung cancer related gene identification that has much better performance than the others, 53.81% AUC estimate in the Sherlock, and 55.66% AUC estimate in the MetaRanker 2.0.

4. Conclusions

In this paper we proposed an integrative strategy of gene prioritization which can employ various genetic resources, including gene expression data, SNP genotype data, and eQTL data, even if it is not limited to use these data only. Particularly, for the integration of scores from different resources, we used the improved TOPSIS method and could make good use of potential complementarities among available genetic resources. To verify the performance of our proposed strategy, we conducted experiments with two datasets regarding prostate cancer and lung cancer, each of which includes gene expression data, SNP genotype data, and eQTL data. The

results demonstrate that our integrative strategy with the improved TOPSIS is superior to other integration methods in combining scores from distinct genetic resources, leading to the better performance in discovering disease-related genes. In addition, compared to other existing gene prioritization tools, our strategy is easily extensible and customizable to use many other resources for the meta-analysis, while producing very impressive results of gene prioritization.

To extend the present work, we are currently developing a web-based application to implement the proposed strategy. The first test version can be found at <http://155.230.107.81/meta.analysis/>.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research was supported by Kyungpook National University research fund, 2012.

References

[1] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

- [2] W. G. Cochran, "Some methods for strengthening the common χ^2 tests," *Biometrics*, vol. 10, pp. 417–451, 1954.
- [3] P. Armitage, "Tests for linear trends in proportions and frequencies," *Biometrics*, vol. 11, no. 3, pp. 375–386, 1955.
- [4] N. H. Chapman and E. M. Wijsman, "Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility," *The American Journal of Human Genetics*, vol. 63, no. 6, pp. 1872–1885, 1998.
- [5] G. Zheng, B. Freidlin, and J. L. Gastwirth, "Robust genomic control for association studies," *The American Journal of Human Genetics*, vol. 78, no. 2, pp. 350–356, 2006.
- [6] P. D. Sasieni, "From genotypes to genes: doubling the sample size," *Biometrics*, vol. 53, no. 4, pp. 1253–1261, 1997.
- [7] T. A. Manolio, "Genomewide association studies and assessment of the risk of disease," *The New England Journal of Medicine*, vol. 363, no. 2, pp. 166–176, 2010.
- [8] T. A. Pearson and T. A. Manolio, "How to interpret a genome-wide association study," *Journal of the American Medical Association*, vol. 299, no. 11, pp. 1335–1344, 2008.
- [9] B. Lehne, C. M. Lewis, and T. Schlitt, "From SNPs to genes: disease association at the gene level," *PLoS ONE*, vol. 6, no. 6, Article ID e20133, 2011.
- [10] L. Consoli, A. Lefèvre, M. Zivy, D. de Vienne, and C. Damerval, "QTL analysis of proteome and transcriptome variations for dissecting the genetic architecture of complex traits in maize," *Plant Molecular Biology*, vol. 48, no. 5–6, pp. 575–581, 2002.
- [11] L. Li, X. Zhang, and H. Zhao, "eQTL," in *Quantitative Trait Loci (QTL)*, S. A. Rifkin, Ed., pp. 265–279, Humana Press, 2012.
- [12] M. F. Moffatt, M. Kabesch, L. Liang et al., "Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma," *Nature*, vol. 448, no. 7152, pp. 470–473, 2007.
- [13] X. He, C. K. Fuller, Y. Song et al., "Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS," *The American Journal of Human Genetics*, vol. 92, no. 5, pp. 667–680, 2013.
- [14] J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga, "ToppGene Suite for gene list enrichment analysis and candidate gene prioritization," *Nucleic Acids Research*, vol. 37, no. 2, pp. W305–W311, 2009.
- [15] L.-C. Tranchevent, F. B. Capdevila, D. Nitsch, B. de Moor, P. de Causmaecker, and Y. Moreau, "A guide to web tools to prioritize candidate genes," *Briefings in Bioinformatics*, vol. 12, no. 1, Article ID bbq007, pp. 22–32, 2011.
- [16] T. H. Pers, P. Dworzynski, C. E. Thomas, K. Lage, and S. Brunak, "MetaRanker 2.0: a web server for prioritization of genetic variation data," *Nucleic Acids Research*, vol. 41, no. W1, pp. W104–W108, 2013.
- [17] C. Li and C. Ye, "Comprehensive evaluation of the operating performance for commercial banks in China based on improved TOPSIS," in *Proceedings of the International Conference on Global Economy, Commerce and Service Science (GECSS '14)*, 2014.
- [18] C.-L. Hwang and K. Yoon, *Multiple Attribute Decision Making: Methods and Applications: A State-of-the-Art Survey*, Springer, 1981.
- [19] R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk, "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments," *FEBS Letters*, vol. 573, no. 1–3, pp. 83–92, 2004.
- [20] R. A. Fisher, *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh, London, 1925.
- [21] R. Walker, M. Thompson, and R. Lawn, *Proficiency Testing in Analytical Chemistry*, Royal Society of Chemistry, 1997.
- [22] U. R. Chandran, C. Q. Ma, R. Dhir et al., "Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process," *BMC Cancer*, vol. 7, article 64, 2007.
- [23] K.-Q. Liu, Z.-P. Liu, J.-K. Hao, L. Chen, and X.-M. Zhao, "Identifying dysregulated pathways in cancers from pathway interaction networks," *BMC Bioinformatics*, vol. 13, no. 1, article 126, 2012.
- [24] X. Mao, Y. Yu, L. K. Boyd et al., "Distinct genomic alterations in prostate cancers in Chinese and Western populations suggest alternative pathways of prostate carcinogenesis," *Cancer Research*, vol. 70, no. 13, pp. 5207–5212, 2010.
- [25] T.-P. Lu, M.-H. Tsai, J.-M. Lee et al., "Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women," *Cancer Epidemiology Biomarkers and Prevention*, vol. 19, no. 10, pp. 2590–2597, 2010.
- [26] T.-P. Lu, L.-C. Lai, M.-H. Tsai et al., "Integrated analyses of copy number variations and gene expression in lung adenocarcinoma," *PLoS ONE*, vol. 6, no. 9, Article ID e24829, 2011.
- [27] E. R. Gamazon, W. Zhang, A. Konkashbaev et al., "SCAN: SNP and copy number annotation," *Bioinformatics*, vol. 26, no. 2, pp. 259–262, 2010.
- [28] R. A. Irizarry, B. Hobbs, F. Collin et al., "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.
- [29] S. Duan, R. S. Huang, W. Zhang et al., "Genetic architecture of transcript-level variation in humans," *The American Journal of Human Genetics*, vol. 82, no. 5, pp. 1101–1113, 2008.